GLOTTALIMAGEEXPLORER – AN OPEN SOURCE TOOL FOR GLOTTIS SEGMENTATION IN ENDOSCOPIC HIGH-SPEED VIDEOS OF THE VOCAL FOLDS

Peter Birkholz

Institute of Acoustics and Speech Communication, TU Dresden, Germany
peter.birkholz@tu-dresden.de

Abstract: This paper introduces the free and open source software GlottalImage-Explorer version 1.0. It is a tool for the efficient, semi-automatic segmentation of the glottis in the frames of high-speed endoscopic videos of the larynx. The implemented segmentation method is based on seeded region growing as described by Lohscheller et al. [7] and has been combined with an intuitive and extensible user interface. The glottal area waveform and the glottis contours can be exported as text files for further processing using other tools.

1 Introduction

The vibratory pattern of the vocal folds determines the sound of our voice. Hence, the analysis of these vibrations is very useful for the assessment of voice pathologies as well as for research in speech production and speech simulation. For example, with regard to the simulation of the voice source, natural oscillation patterns can be used to evaluate self-oscillating computer models of the vocal folds [2, 3].

High-speed laryngoscopy with frame rates as high as 4000 frames/s nowadays allows a very detailed analysis of vocal fold vibration, because it can resolve each individual oscillation period in multiple images. However, the amount of data captured with these cameras is very high and demands semi-automatic or fully automatic methods for analysis. A feature of paramount interest in the frames is the shape of the glottis, i. e., the glottis contour. From the sequence of glottis shapes in the frames, other features or waveforms of interest can be derived, for example the glottal area waveform, the glottovibrogram [6], or the kymogram [10].

The robust extraction of the glottis in endoscopic images is a challenging task, and many methods have been proposed. Existing methods apply, for example, region growing [11, 7], active contours and the watershed transform [6, 1], or a region-based level-set strategy [4].

So far, the existing implementations for glottis segmentation are either closed source or not freely available. This makes it hard to explore improvements of existing approaches by anyone but the developers and to compare different approaches with one another. Therefore, we present in this paper the first free and open-source implementation of a glottis segmentation method, which is based on the seeded region growing method by Lohscheller et al. [7]. This method was selected, because it is conceptually simple and clinically evaluated. However, compared to more recent fully automated methods, it requires a small amount of user interaction. The software is called GlottalImageExplorer (GIE) and available for download from `http://www.vocaltractlab.de`.

2 Method for glottis segmentation

In the following we give a brief overview of the glottis segmentation method, which is based on seeded region growing and described in more detail in [7]. The method works with grayscale images (color images are automatically converted to grayscale images for segmentation) and exploits the fact that the glottis is usually darker than the surrounding tissue. Assume that the position of one or more (seed) points, i. e., pixels, within the glottal area is known. These pixels constitute the initial part of the segmented region. Then the algorithm evaluates the gray values of the pixels in the direct neighborhood of the seed points and adds them to the region, if their gray values are lower (darker) than a specific threshold η. This process is iterated on with the pixels neighboring the region until there are no new pixels that satisfy the threshold criterion.

Due to gradients in the illumination of the glottis, this method would not work well if there was only a single fixed threshold for the whole image. Instead, we assume an individual threshold for each row of pixels, i. e., a threshold progression $\eta(y)$, where y $(0 \leq y < Y)$ is the row index and Y is the number of rows. Since the illumination conditions vary slowly with y, it is sufficient to manually specify $\eta(y)$ at a few rows, and linearly interpolate between them for the remaining rows. A manually specified set of seed points in combination with a threshold progression based on a few manually specified threshold values is sufficient for the accurate segmentation of the glottis. In our software, the user can specify the position of three seed points and three threshold points. To further simplify the manual adjustments, the seed points can be coupled to the threshold points, meaning that $\eta(y)$ is always defined at the y-coordinates of the three seed points (and automatically interpolated in-between).

Given that the illumination conditions may also vary over time, a single threshold progression $\eta(y)$ is usually not sufficient for all frames in a film. Similarly, the glottis position may slightly move over time, so that a single set of seed points may not be appropriate for all frames. Therefore, the user may specify the seed points and the threshold progression in any selected frame. For the frames in-between, both the seed points and the threshold progression is interpolated (over time). Usually, the manual specification in a handful of frames is sufficient for satisfactory segmentation results in a whole sequence of, e. g., 4000 frames.

In the current version of GIE, the size of the frames is fixed to 256x256 pixels and the number of frames in a film is limited to 10,000.

3 Graphical user interface

This section gives a brief overview of the graphical user interface of the software and is also contained in the user manual. Figure 1 shows the graphical user interface, which consists of a menu, a control panel at the left side, three images in the top-right area, and the glottal area waveform at the bottom.

3.1 Menu

The File menu allows loading a film from an AVI file (*.avi), a raw data file of the HRES ENDOCAM system of the company Richard Wolf GmbH (*.bld), or from a set of BMP images (*.bmp) in a folder. The image size in the AVI and BMP files must be 256x256 pixels.

Furthermore, there are two menu items to load and save the segmentation data (*.seg). The segmentation data are the user-specified threshold and seed point settings. They are saved in a simple text format, where each line represents the seed points and threshold points of one frame.

Finally, there are menu items to export the glottal area waveform and the glottis contour waveform, both as text files (*.txt). The text file for the glottal area waveform simply consists of one

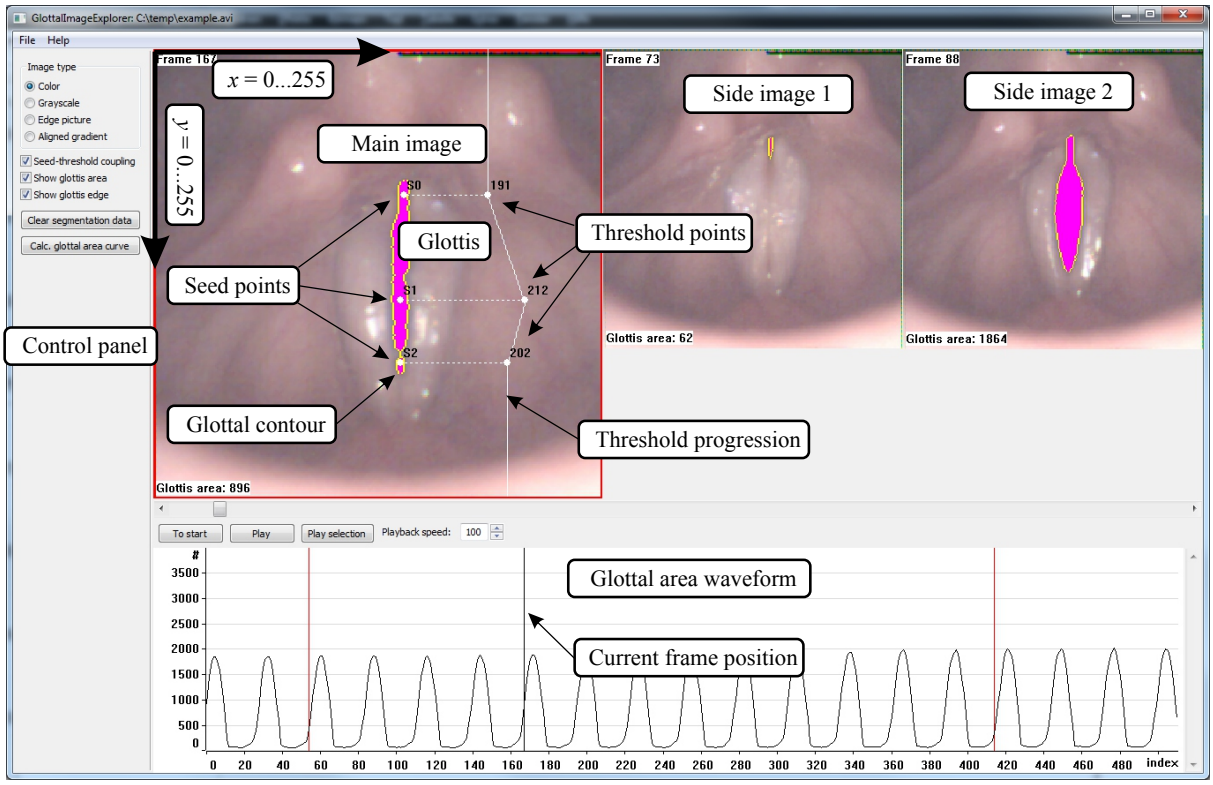

Figure 1 - Graphical user interface of GlottalImageExplorer.

glottal area value (number of pixels segmented as the glottis) for each frame in the film. The text file for the glottis contour waveform contains two text lines for each frame that specify the left and right part of the glottis contour. Each contour consists of the x-values of the pixels in the pixel rows 0 (top) to 255 (bottom). For the pixel rows above and below the glottis, the x-values of the contours in the file are set to zero. The glottal contours written to the file are meant for further processing or plotting in other programs like MATLAB.

3.2 Control panel

The upper part of the control panel contains radio buttons to specify the way the frames of the film are displayed: either as color images (original data), grayscale images, edge pictures or "aligned gradient" pictures. The grayscale images are used for the actual glottis segmentation, i. e., as basis for the region growing method. The edge and gradient pictures emphasize the edges in an image, applying the Sobel operator [5]. This can help the user adjust the region growing thresholds such that the contours of the segmented glottis correspond well to the real edges of the glottis.

Below the radio buttons there are three checkboxes. When the checkbox "Seed-threshold coupling" is checked, the y-coordinates of the threshold points automatically assume the y-coordinates of the seed points, i. e., they cannot be changed independently. In this case, the threshold values for the region growing algorithms are always specified at the vertical positions of the seed points. When the checkbox "Show glottis area" is checked, the segmented glottis area is drawn in purple color in the images. When the checkbox "Show glottis edge" is checked, the contour of the segmented glottis is drawn in yellow.

Below the checkboxes are the buttons "Clear segmentation data" and "Calc. glottal area curve". The first button clears all user-defined adjustments of seed points and threshold points. The second button (re-)calculates the glottal area waveform based on the current seeded region growing

settings. After the calculation, the area waveform is shown in the bottom part of the program window.

3.3 Images

Three frames of the film are shown as images in the main part of the program window. The left (main) image shows the *current* frame, where the user can modify the position of the seed points and the threshold points. The two images at the right side (side images) show two frames that can be selected by the user to facilitate visual comparisons of frames at different positions in the film. The current frame can be selected to be shown in one of the side images with the context menu (right-click) in the main image.

The current frame shown to the left can be set using the scrollbar below the images or changed to the previous or the next frame with Ctrl + ← or Ctrl + →. For every frame the user can choose to specify the segmentation data manually or let the software interpolate these data between the neighboring frames with manually specified data. For the current frame, manual segmentation is selected with the item "User defined segmentation" in the context menu of the main image. If "User defined segmentation" is selected, the main image gets a red border and the seed points and threshold points can be dragged with the mouse. The three seed points are labeled with "S0", "S1", and "S2", and must be moved into the area of the glottis. The software restricts the movement such that S0 is always the uppermost and S2 the lowermost of the three points.

There are three threshold points that define the threshold progression for the seeded region growing from the top to the bottom of the frame. Each of these points specifies the threshold value at a certain y-coordinate of the frame. The threshold value associated with a point is represented by its x-coordinate, i. e., the further left a threshold point, the lower the threshold value, and vice-versa. The threshold values are displayed next to the points. The threshold progression is defined by the linear interpolation between the threshold points and indicated as a white line in the image. If the threshold points are coupled to the seed points, only the threshold *values* can be changed (x-coordinate), but their y-coordinates are forced equal to the seed points' y-coordinates. In this case, seed points and threshold points are connected by dashed horizontal lines. If selected in the control panel, the segmented region is shown as a purple area and the glottis contour is shown in yellow.

Below the scrollbar for the current frame are buttons to play the sequence of frames as a movie. The button "Play" starts to play the sequence from the current frame on, and the button "To start" sets the first frame as the current frame. The "Playback speed" can be set to any number between 1 and 100, where 1 corresponds to very slow playback and 100 to the fastest possible playback of the frame sequence. The actual number of frames per seconds depends on the speed of the computer.

3.4 Area waveform

The glottal area waveform shows the time function of the glottal area according to the segmentation for a certain range of frames around the current frame. The calculation of the waveform must be manually triggered with the button "Calc. glottal area curve" in the control panel, because the calculation takes a couple of seconds. The current frame is indicated by a black vertical line and the frames with user-defined segmentation data are shown as red vertical lines. Using the context menu of the area waveform panel, the waveform can be zoomed in and out and the amplitude can be scaled.

4 Segmentation examples

Figure 2 shows examples of segmented glottal area waveforms obtained from two endoscopic recordings of the same male subject. The recordings were made with an ENDOCAM camera system of the company Richard Wolf GmbH (Knittlingen, Germany) using a rigid endoscope with 4000 frames/s and durations of 2 seconds. The manual adjustments in GIE that were necessary to obtain the area waveforms (including visual double-checking) took about 15 min for each of the two films. For the first recording, the subject sustained the vowel /ɛ/ with a breathy voice (upper panel), and for the second recording he sustained the same vowel with a modal voice. The waveforms nicely reflect the expected differences between the two phonation types [8]. Breathy voice area pulses have a higher amplitude (both AC and DC) than the modal voice area pulses, and an offset due to incomplete closure in terms of a permanent glottal chink in the posterior part of the glottis. Furthermore, the breathy voice pulses are slightly skewed to the left, while the modal voice pulses are almost symmetric.

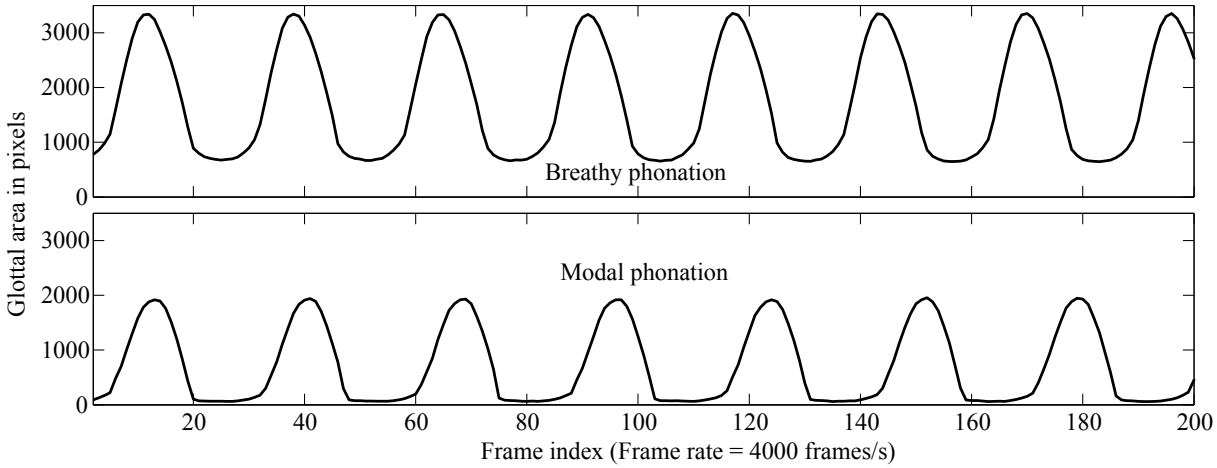

Figure 2 - Examples of glottal area waveforms for breathy phonation (top) and modal phonation (bottom) extracted with GlotalImageExplorer.

5 Concluding remarks

In summary, the software allows the fast segmentation of the glottis in sequences of frames with quite little adjustments necessary by the user. Currently, the software assumes images with a fixed resolution of 256x256 pixels. However, this can be easily changed in the source code on demand. The software also provides a platform for the future implementation of alternative, possibly fully automated segmentation methods, as [4, 1]. Beyond that, it can be extended with more elaborate postprocessing methods, for example with an algorithm to directly extract the glottovibrogram [6]. GIE could also serve as a platform to extract interesting features other than the glottal shape, for example the propagation of the mucosal waves during phonation as proposed in [9], or the movement of arytenoids.

Acknowledgments

The author thanks Alexander Mainka for the acquisition of the endoscopic high-speed videos that were analyzed for the area waveforms in Figure 2.

References

[1] G. Andrade-Miranda, J. I. Godino-Llorente, L. Moro-Velázquez, and J. A. Gomez-Garcia. An automatic method to detect and track the glottal gap from high speed videoendoscopic images. *Biomedical Engineering Online*, 14(1):100, 2015.

[2] P. Birkholz. A survey of self-oscillating lumped-element models of the vocal folds. In B. J. Kröger and P. Birkholz, editors, *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2011*, pages 47–58. TUDPress, Dresden, 2011.

[3] B. D. Erath, M. Zañartu, K. C. Stewart, M. W. Plesniak, D. E. Sommer, and S. D. Peterson. A review of lumped-element models of voiced speech. *Speech Communication*, 55(5):667–690, 2013.

[4] O. Gloger, B. Lehnert, A. Schrade, and H. Volzke. Fully automated glottis segmentation in endoscopic videos using local color and shape features of glottal regions. *IEEE Transactions on Biomedical Engineering*, 62(3):795–806, 2015.

[5] R. Gonzalez and R. Woods. *Digital Image Processing*. Addison Wesley, 1992.

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

[7] J. Lohscheller, H. Toy, F. Rosanowski, U. Eysholdt, and M. Döllinger. Clinically evaluated procedure for the reconstruction of vocal fold vibrations from endoscopic digital high-speed videos. *Medical Image Analysis*, 11(4):400–413, 2007.

[8] H. Pulakka, P. Alku, S. Granqvist, S. Hertegard, H. Larsson, A.-M. Laukkanen, P.-A. Lindestad, and E. Vilkman. Analysis of the voice source in different phonation types: simultaneous high-speed imaging of the vocal fold vibration and glottal inverse filtering. In *INTERSPEECH-2004*, pages 1121–1124, Jeju Island, Korea, 2004.

[9] D. Voigt, M. Döllinger, U. Eysholdt, A. Yang, E. Gürlek, and J. Lohscheller. Objective detection and quantification of mucosal wave propagation. *Journal of the Acoustical Society of America*, 128(5):EL347–EL353, 2010.

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

[11] Y. Yan, X. Chen, and D. Bless. Automatic tracing of vocal-fold motion from high-speed digital images. *IEEE Transactions on Biomedical Engineering*, 53(7):1394–1400, 2006.