Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication xxx (2011) xxx–xxx

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Automatic glottal segmentation using local-based active contours and application to glottovibrography

Sevasti-Zoi Karakozoglou[a,c,d,*], Nathalie Henrich[b], Christophe d'Alessandro[a], Yannis Stylianou[c]

[a] LIMSI-CNRS, Orsay Cedex, France
[b] Department of Speech and Cognition, GIPSA-lab, UMR 5216: CNRS, INPG, University Stendhal, UJF, Grenoble, France
[c] Computer Science Department, University of Crete, Heraklion, Greece
[d] Computer Science Department, University Paris-Sud 11, Orsay Cedex, France

Abstract

The use of high-speed videoendoscopy (HSV) for the assessment of vocal-fold vibrations dictates the development of efficient techniques for glottal image segmentation. We present a new glottal segmentation method using a local-based active contour framework. The use of local-based features and the exploitation of the vibratory pattern allows for dealing effectively with image noise and cases where the glottal area consists of multiple regions. A scheme for precise glottis localization is introduced, which facilitates the segmentation procedure. The method has been tested on a database of 60 HSV recordings. Comparisons with manual verification resulted in less than 1% difference on the average glottal area. These errors mainly come from detection failure in the posterior or anterior parts of the glottal area. Comparisons with automatic threshold-based glottal detection point out the necessity of complete frameworks for automatic detection. The glottovibrogram (GVG), a representation of glottal vibration is also presented. This easily readable representation depicts the time-varying distance of the vocal-fold edges.

Keywords: High-speed videoendoscopy; Vocal-fold vibration; Active contours; Representation; Glottovibrogram; Electroglottography

1. Introduction

High-speed videoendoscopy (HSV) is the most promising approach to directly assess vocal-fold vibrations. Despite this, its application to clinics is limited by the vast amount of data that must be evaluated both qualitatively and quantitatively (Deliyski and Petrushev, 2003; Deliyski et al., 2008). There is a need to reduce the dimensionality of the spatio-temporal information, and to efficiently represent the high-speed data in a compact, easy to use and lossless way. In this regard, automatic segmentation of the glottal area is a major challenge. Recently, several methods have been proposed for glottal segmentation using two different segmentation approaches; either region-based or model-based ones.

The most straightforward region-based methods use a threshold or histogram approach (Kohler, 1981; Haralick and Shapiro, 1985). Peaks and valleys in the histogram of an image, computed from pixel color or intensity information, are used to classify clusters within the image. The histogram is assumed to be at least bimodal (two peaks), which is not the case in images with low contrast and objects with heterogeneous profiles. The most recent use of such an approach is found in Mehta et al. (2010), Mehta et al. (2011). The method, however, is not automatic as it requires visual inspection and threshold adjustments.

The second main group of region-based methods are seeded region-growing methods (SRG). SRG methods examine neighboring pixels of an initial set of seed points

[*] Corresponding author at: LIMSI-CNRS, Orsay Cedex, France.
E-mail addresses: skarako@csd.uoc.gr (S.Z. Karakozoglou), Nathalie.Henrich@gipsa-lab.grenoble-inp.fr (N. Henrich), Christophe.D'Alessandro@limsi.fr (C. d'Alessandro), yannis@csd.uoc.gr (Y. Stylianou).

0167-6393/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2011.07.010

and determine whether the neighboring pixels should be added to the region (Adams and Bischof, 1994; Mehnert and Jackway, 1997). Such a method requires robust criteria and relatively clear edges in order to converge to the region of interest. A SRG method was proposed by Yan et al. (2006) for glottis segmentation. An initial region of interest has to be defined manually and the seed points are computed by assuming a Rayleigh distribution of intensity. The proposed method neither takes advantage of the vibration pattern nor considers special constraints for frames depicting closed glottis. A supervised SRG method was presented by Lohscheller et al. (2007). In this study, seed points are defined by the user in a single selected image. The use of a two-dimensional threshold matrix is proposed as the stopping criteria. The seeding procedure is reiterated within the glottal cycle intervals. The segmentation is supervised and certain parameters are chosen during the process. A recent SRG-based method was developed by Demeyer et al. (2009). The seed point for the region-growing method is defined as the maximal response of a laplacian of a gaussian filter. Intensity is used as the sole homogeneity criterion. Size is uncertain and can be underestimated. This method is applied to periodic frames, where the glottis is supposed to be at maximum opening. The region-growing results are propagated to the rest of the sequence using a level-set method. Parameters are chosen empirically.

Model-based methods, such as active contours, make use of the idea that objects of interest have some kind of repetitive form of representation. Active contours, also known as snakes, are mainly used to dynamically locate the contour of an object. A snake is an energy-minimizing spline guided by external constraint forces and influenced by image forces that pull it towards desired features, such as lines and edges (Kass et al., 1988). By properly choosing an initial contour near the object of interest, the model will converge to the desired solution. An energy function is associated with the curve in terms of its shape and distance from desired image features. The problem of object detection is thus treated as an energy minimization problem. Marendic et al. (2001) were the first to present an active contour algorithm for vocal-fold extraction from high-speed video data. The vocal-fold vibratory pattern is taken into consideration to reinitialize the algorithm. They applied their algorithm in one sequence, using empirically chosen parameters. Allin et al. (2004) used a snake-based segmentation for the medial edges of the vocal folds. Although they used data from a low-speed stroboscopic system, their approach is interesting because they use the Fischer linear discriminant to achieve a coarse segmentation of the sequence. This method demands the training of a color classifier for each sequence from more than one frame. Then, active contours are used to refine the result. Lohscheller et al. (2004) used a combination of threshold technique and active contour to segment the pseudoglottis. The most recently developed method that uses active contours is presented in Moukalled et al. (2009). They employed a pair of open-curve snakes on the digital kymographic sequences. This method requires the user to define the posterior and anterior points in an image and to verify the segmentation result of one DKG frame, before it propagates to the rest of the sequence. The segmentation results are applied to the HSV sequences once the segmentation is completed.

The active contour based approach seems appropriate for the purpose of automatic segmentation with no user intervention. Recent refinements of the image processing technique can be applied to improve the dynamic glottal-edge detection. We propose to explore the applicability of the local region-based framework which has been proposed by Lankton and Tannenbaum (2008). This method allows the foreground and background to be modelled in terms of smaller local regions, instead of representing them with global statistics. This allows us to deal with inhomogeneity, common in medical images. The energy function is computed locally and energy minimization is performed by fitting a model to each local region.

The segmented glottal area of the entire high-speed sequence can be transformed into a two-dimensional representation. The first attempt of glottal shape representation was made by Westphal and Childers (1983). The Phonovibrogram representation (PVG) presented by Lohscheller et al. (2008) transforms vocal-fold movements into geometric objects. PVG requires one to calculate the distance of the edges to the glottal axis, a method very sensitive to the accuracy of glottal-axis detection. Therefore, we present a recently proposed approach to Lohscheller's two-dimensional representation of the glottal area's shape, the Glottovibrogram (GVG) (Karakozoglou, 2010; Einig, 2010; Döllinger et al., 2011).

A new method for fully automatic detection of glottal edges is described in Section 2. In Section 3 the GVG representation is presented. The database and evaluation procedure are given in Section 4. Section 5 presents the segmentation evaluation and applications to glottovibrography. Finally, Section 6 concludes this work.

2. Algorithm for glottal detection

The proposed method is described in the present section and depicted in Fig. 1. The procedure is typical for segmentation without user intervention, as it has also been presented by Demeyer et al. (2009).

2.1. Glottis localization and landmark frames

In a first step, the algorithm selects the region of interest and extracts useful information about the glottal cycles.

The open glottis is the darkest region in an image and the one that varies the most in time. Within each glottal cycle, the frame with maximal glottal opening can be detected as the one for which the sum of pixel intensities is minimum. Such a frame is labeled as a landmark frame

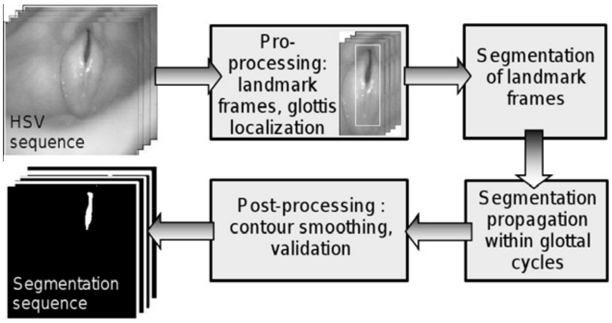

HSV sequence
Pro-processing: landmark frames, glottis localization
Segmentation of landmark frames
Segmentation propagation within glottal cycles
Post-processing : contour smoothing, validation
Segmentation sequence

Fig. 1. Outline of the glottal detection algorithm.

(Eq. (1)). The landmark frames represent the maximal open states of the glottal cycle within the sequence under consideration. The same intuition for frames depicting maximal glottal area has also been used with some variations in previous studies (Demeyer et al., 2009; Lohscheller et al., 2007). To ensure that all selected landmark frames represent maximal glottal areas, the ones with high overall intensities are checked and those that correspond to high overall mean intensities are supressed.

$$I_{landmark} = \underset{i=1..k}{\operatorname{argmin}} \left(\sum_{x} \sum_{y} I_i(x,y) \right) \quad (1)$$

The region of interest covers only a part of the entire image. For localization and computational reasons, there is no need to process the entire image. The image size of the high-speed sequences is 256 × 256 pixels. The glottal area, and so the region of interest, usually covers less than 25% of the entire image size in the present database. However, camera tuning for the same purpose is similar, since recording usually involves the entirety of the larynx for multiple studies, for both medical doctors and researchers. Edge-based morphological processing of a landmark frame is applied to each landmark frame in order to find a large, nearly vertically oriented area. A Sobel filter is used to detect strong edges in the vertical direction. A morphological closing is then performed on the gradient map, so as to connect small related regions. These regions are detected by connected component analysis (Samet and Tamminen, 1988). The object with the largest area and vertical orientation is selected. A rectangular area surrounding the selected area is computed, termed the bounding box. This step allows for a reduction in the amount of data to be processed and treatment of larger video sequences. The coordinates of the cropped rectangle are stored, and they are used to apply the segmentation result to the initial sequence. An example is presented in Fig. 2(a). To locate the glottis with higher accuracy, the same edge-based processing is applied to get a tighter bounding box surrounding the glottal region in each landmark frame. The bounding box remains steady within each glottal cycle in order to compensate for glottal drift and/or movements of the endoscope (see Fig. 2(b)).

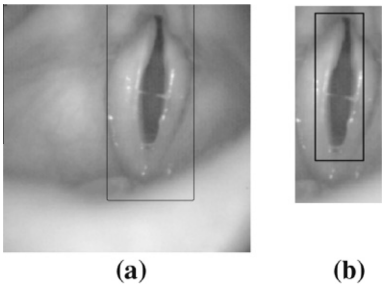


Fig. 2. Size reduction by glottis localization. (a) The bounding box on the original image (256 × 256 pixels). (b) The cropped image (202 × 89 pixels) and the bounding box surrounding the glottal area. The original sequence is reconstructed from the cropped video sequence without any loss.

It is also necessary to determine whether the glottis exists in all images or not. In an active contour framework, the algorithm may evolve into excrescences if further constraints are not applied. This is solved with two techniques. First, the pixel intensities of each image are taken into account. If the minimum pixel intensity in the image persists over a global threshold, defined as the median of the pixels intensities of the entire sequence, a glottal opening is assumed to be absent. Second, the bounding box is computed. If it is centered far from the bounding box of the landmark frame, or if it does not exist, it is assumed that there is no glottal opening. When both of these conditions occur simultaneously, the frame is excluded from further processing.

Simple contrast enhancement is also performed on the high-speed video sequences. In order to improve the local contrast in the images, bringing out more detail in the glottal area while avoiding significant noise introduction, the contrast limited adaptive histogram equalization algorithm (CLAHE) was used (Zuiderveld, 1994). It consists of a generalized version of adaptive histogram equalization that computes several histograms, each corresponding to a distinct section of the image, which are then used to redistribute the lightness values of the image, thus compensating for noise amplification.

The enhanced video sequence is used for the following steps of the algorithm. The automatic active-contours framework dictates the knowledge of landmark frames, as well as information about the area and shape of the object of interest. Each landmark frame is therefore used as a reference for the segmentation propagation within each glottal cycle, as explained in the following sections.

2.2. Glottis segmentation: main principles

The segmentation method used in this work is based on the framework proposed by Lankton and Tannenbaum (2008), referred to as local region-based framework for guiding active contours. The approach models the foreground and background in terms of smaller local regions, since foreground and background regions cannot always be represented with global statistics. This framework provides correct conversion in instances of inhomogeneity which are common in medical images. In the present method, the constant intensity Chan–Vese model is used (Chan and Vese, 2001). This approach models the foreground and background as constant intensities represented by their means. Mean intensities of exterior (v) and interior (u) regions are computed in proximity of the curve (locality defined by $\mathcal{H}\phi(y)$, where $\mathcal{H}$ is the approximation of the smooth Heavyside function and $\phi(y)$ is the level set function Sethian, 1999), allowing us to ignore any inhomogeneities distant from the glottal area (Eq. (2)). Furthermore, the use of local information allows the curve to split or merge, using only the contribution of the neighbourhod statistics for a given point $I(x,y)$, where x, y are independent spatial coordinates.

$$E = \int_{\Omega_y} \left(\mathcal{H}\phi(y)(I(y) - u)^2 + (1 - \mathcal{H}(y))(I(y) - v)^2\right)\mathrm{d}y \tag{2}$$

The size of local neighborhoods is defined by the size of the bounding box for each frame; the size is equal to 1/3 of the smallest dimension of the bounding box. Due to inhomogeneities which may occur in the image, the local region is restricted to the smallest possible size so as to ensure maximal separability. The above parameters are applied to every frame of the sequence.

The initial contour curve is of major importance in active contour methods. Correct curve placement facilitates the segmentation procedure and points out the object of interest. The initial mask for the landmark frames is therefore chosen differently than for the rest of the sequence. The initial mask of landmark frames must be computed. For the remaining sequences, the final contour from the previous treated frame is used. Once the initial mask for a frame is chosen, the algorithm runs until convergence; either until no changes are observed in the contour, or 150 iterations are reached.

The aforementioned segmentation procedure is applied to each frame of the HSV sequence with the same parameters. In order to automatically segment the entire sequence, the framework dictates the initialization of the segmentation procedure at the landmark frames of each glottal cycle, as presented in the following section.

2.3. Segmentation of landmark frames

The segmentation algorithm starts from the landmark frames of a sequence. The glottis is an object with a heterogeneous feature profile. Even though it is darker than the surrounding tissues, there might be regions where local statistics do not provide substantial similarity criteria. As the initial contour curve is crucial in active contour methods, we propose the use of two automatic methods for curve initialization on landmark frames.

Algorithm 1. Pseudocode for Segmentation of Landmark Frames. T_o refers to the total number of Landmark frames. The function ActiveContours refers to the algorithm suggested by Lankton and Tannenbaum (2008)

```
Input: Landmark Frames, Bounding Boxes
Output: Segmentation of Landmark Frames
foreach Landmark Frame $k$, $k = 1 : To$ do
    Calculate InitialMapT of Frame $k$ as ThresholdingMap ;
    Calculate InitialMapL of Frame $k$ as LocalizedMap ;
    FinalMapT = ActiveContours(Frame $k$, InitialMapT) ;
    FinalMapL = ActiveContours(Frame $k$, InitialMapL) ;
    if BestFits(FinalMapT) > BestFits(FinalMapL) then
        Segmentation of Frame $k \cong$ FinalMapT ;
    else
        Segmentation of Frame $k \cong$ FinalMapL ;
    end
end
```

Algorithm 1 depicts the segmentation of landmark frames. The initial contour is estimated with two methods. The first method consists of finding the intensity threshold within the bounding box of each landmark frame. In cases of high contrast, the glottal region is much darker and relatively homogeneous so that a threshold is sufficient for initial discrimination. The threshold is found by selecting the minimum of a smoothed bimodal histogram. This is also referred as the mode method (Glasbey, 1993). However, the assumption of the bimodal histogram is not always valid, as it depends on the statistics of the image. To address this problem we suggest the use of a localization-based map (second method). For this purpose, an ellipse is computed, whose center is located on the center of the bounding box and its size is proportional to the bounding box's size. Its orientation is based on the orientation of the glottis computed during the localization. This ellipse-shaped mask covers the glottal area and points out with good accuracy where the active contours should converge. The computation of the final contour is performed as described in Section 2.2. Comparison of the computed contours using the above two methods is based on the orientation and maximal separability of the object relative to the background in terms of intensity. The computed contour which best fits the above criteria is then used as the segmentation mask of the frame and also for the propagation of the segmentation to the remaining elements of the sequence. The segmentation propagation procedure is presented in the following section. Fig. 3 provides an example of comparison between the two methods for the segmentation of landmark frames. In Fig. 4 an example of segmentation on a single landmark frame is shown, beginning from the ellipse-shaped initial mask.

Please cite this article in press as: Karakozoglou, S.Z. et al., Automatic glottal segmentation using local-based active contours and application to glottovibrography, Speech Comm. (2011), doi:10.1016/j.specom.2011.07.010

2.4. Propagation of segmentation

Once all landmark frames have been segmented, the remaining frames are processed. To ensure temporal consistency, the segmentation is propagated by using as initial mask for the kth frame the segmentation result from the $(k-1)$th (forward) or the $(k+1)$th frame (backward), depending the position of the landmark frame (Algorithm 2).

Algorithm 2. Pseudocode for Segmentation Propagation

```
Input: HS-sequence, Landmark Frames, Final Contours of Landmark
       Frames
Output: Segmentation matrix of HS-sequence
foreach Landmark Frame $k$, $k = 1 : To$ do
    foreach Frame within the glottal cycle
    $m = T_o(k) + 1 : +1 : T_o(k+1) - 1$ do
        Calculate FinalContour of Frame $m$ using FinalContour of
        Frame $m - 1$;
    end
end
```

2.5. Post-processing

There may be cases where the segmentation may converge to inconsistent regions. As such, only contours which are present within the limits of the bounding boxes are retained, in order to suppress undesired contours. To ensure maximal separability in terms of local statistics, the mean intensity of the segmented regions with respect to its surroundings is compared. More specifically, if the mean intensity of a segmented region is significantly higher than the surroundings or other regions within the same image, the region is then excluded as it will most likely have not captured the glottal area. Furthermore, for each segmented object whose histogram is found to be bimodal with a high intensity threshold, a threshold segmentation is then applied, in order to capture the homogeneous region with low intensity. The segmentation matrix was finally smoothed in order to exclude holes in the found regions.

3. Glottovibrograms: a new proposal for data visualization

Dimensionality reduction is the most important aspect of high-speed visualization. Spatio-temporal information must be represented without loss of information. In Lohscheller et al. (2008) the Phonovibrogram (PVG) was introduced, which is a further development of spatio-temporal plots of vocal-fold vibrations (Neubauer et al., 2001). The PVG is a 2-D diagram of vocal-fold vibrations. This representation transforms vocal-fold movements into well-defined geometric objects, thus allowing direct assessment of the vocal-fold dynamics of an entire video sequence in a single image. However, PVG visualizes the deflections of the medial vocal-fold edges from the glottal axis. This representation can be difficult to interpret and strongly depends on the detection of the glottal axis. Therefore the method is adapted to allow for a better visualization of the deflection between the medial vocal-fold edges. This visual representation is termed the Glottovibrogram[1] (GVG). It has been independently proposed recently by two research groups (Karakozoglou, 2010; Einig, 2010; Döllinger et al., 2011).

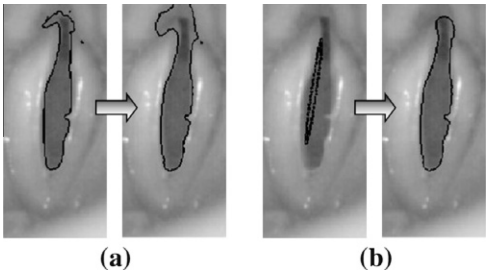


Fig. 3. The process of curve initialization: (a) curve initialization by thresholding; the initial curve and the final curve and (b) curve initialization by ellipse; the initial curve and the final curve. For each landmark frame, the regions defined by segmentation algorithm convergence are compared. The contour that best fits the glottal area is used for propagation.

Instead of measuring the distance between the glottal symmetry axis and the vocal-fold contours, as proposed by Lohscheller et al. (2008), the distance between the vocal-fold contours themselves is calculated, specifically the distance of points found across the glottal axis perpendicular to the glottal axis line (Eq. (3) and Fig. 5). The GVG offers a more representative image of the vibration, even in the presence of artifacts. Glottal axis detection is of major importance and cannot be avoided. The detection of the glottal axis strongly depends on the geometry of the detected glottal area, as the axis is the area's symmetry line. Visualizing the deflection of the vocal-fold edges, instead of their deflection from the glottal axis, provides a more intuitive representation of vocal-fold vibration evolution as would be shown in a HSV sequence. By calculating the distances of the vocal-fold edges, we have managed to depict a well-shaped form of the vocal-fold vibration, even when detection errors occur.

$$\delta^{gl}(m,t) = \|c_L(m,t) - c_R(m,t)\|_2, \quad \forall m \tag{3}$$

The GVG computation is based on the PVG formulation explicitly presented in Lohscheller et al. (2008). For the GVG representation, the vocal-fold edges $c_{L,R}(m,t)$ are equidistantly sampled with $m \in [0, M]$. For each image $I(x,y,t)$, the distances $\delta^{gl}(m,t)$ are computed among points perpendicular to the glottal axis (Eq. (3)). The distances $\delta^{gl}(m,t)$ between the left $c_L(m,t)$ and right $c_R(m,t)$ vocal-fold edges are stored in matrix $\mathcal{D}^{gl}$, which is color coded for visualization (grayscale colormap). The distances are normalized within the interval $[0,1]$, with 0 corresponding to zero distance and 1 corresponding to maximal distance.

[1] The GVG is presently included in the PVGA analyzer software. This appeared to the authors following a personal communication with Prof. Lohscheller at the 9th International Conference on Advances in Quantitative Laryngology, Voice and Speech Research, September 2010.

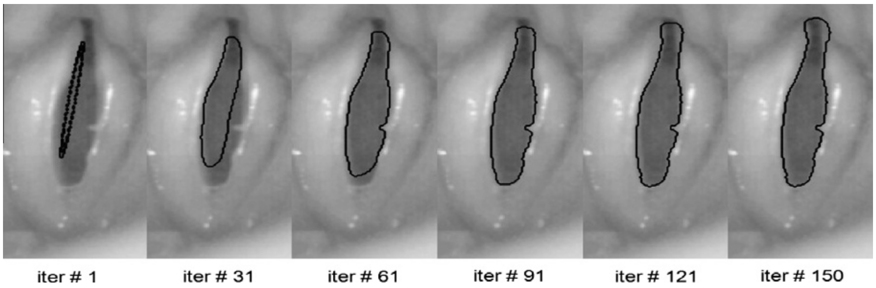


Fig. 4. Curve evolution in a landmark frame. From the ellipse-shaped curve (iteration #1), the curve evolves until it converges to the medial vocal-fold edges (iteration #150). Curves are shown at 30 iteration intervals.

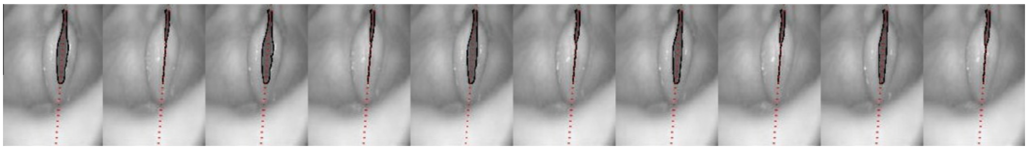


Fig. 5. Linked contours with corresponding glottal axis.

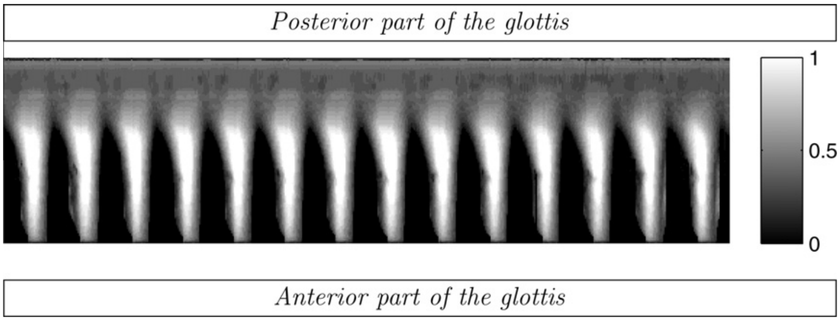


Fig. 6. Glottovibrogram of a high-speed sequence. Black corresponds to zero distance between the vocal-fold edges and white corresponds to the maximum observed distance. Sequence 1; see Appendix.

An example of the GVG representation is shown in Fig. 6. It presents the glottal movement on a 125 ms high-speed sequence, which corresponds to 14 glottal cycles. Time is presented on the x-axis, while distances $\delta^{gl}(m,t)$ between left and right vocal-fold edges are the y-axis. For each cycle, the depicted glottal movement is a zipper-like posterior-to-anterior opening followed by an abrupt closure. The remaining gray area in the top part of Fig. 6 corresponds to a permanent glottal chink, i.e. an absence of glottal closure in the posterior part of the glottis.

Visualization of the velocity pattern along the length of the vocal-folds is interesting. By computing the derivative of the distance, we can depict the speed profile of the vibrations on this representation. This can be done by superimposing the visualization of the derivative of the $\mathcal{D}^{gl}$ matrix on the GVG representation. An example of this joint representation is shown in Fig. 7.

4. Material and methods

4.1. Database

The data used during this work were taken from a high-speed database recorded at the University Medical Center Hamburg-Eppendorf (UKE) in Hamburg, Germany, by the team of Prof. Hess (Frank Müller and Götz Schade) and co-author Dr. Henrich. It consists of synchronized high-speed video, audio, and EGG recordings of two male subjects; one speaker (S1) and one singer (S2), performing different voice qualities, pitches, and transitions.

For the high-speed recordings, a rigid endoscope (Wolf 90 E 60491) equipped with a continuous light source (Wolf 5131) driven by optic fiber was used. The data were recorded at 4000 fps. Along with the high-speed recording, the glottal contact signal was acquired by an electroglottograh (Glottal Enterprises, EL-2 type Rothenberg, 1992). Electroglottography (EGG) is the most common non-invasive technique for measuring variations in vocal-fold contact area by passing a small-intensity high-frequency current between two electrodes secured around the neck at the level of the larynx (Gilbert et al., 1984; Childers and Krishnamurthy, 1985; Scherer et al., 1988). The EGG and audio signals were sampled at 44170 Hz, directly on the medical platform. Real-time monitoring of the EGG signal was performed for each recording with an A/D oscilloscope. The purpose of the experiment was to compare EGG features and glottal behavior in different spoken and sung situations.

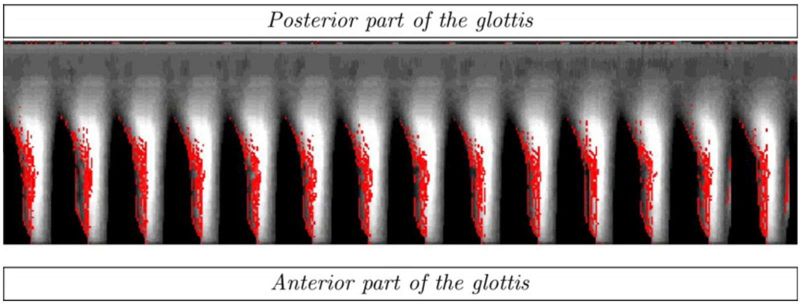


Fig. 7. GVG with maximum speed profile of a video sequence. Red regions correspond to points where the vocal-folds move with maximum velocity. Sequence 1; see Appendix. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

For the purpose of this work, 60 recordings from this database were used. Sequences were chosen from the middle of the recorded sequence. Each sequence contained 501 frames, corresponding to roughly 125 ms.

4.2. Manual segmentation

Automatic segmentation results have been manually verified and corrected if necessary with an interactive tool adapted from (Henrich, 2001; Bailly, 2009). The tool consists of an interface implemented in Matlab which interacts with the given input and treats the contour using Bezier splines (Bezier, 1972). The video and segmentation quality was subjectively evaluated by 13 participants, seven of whom were familiar with voice analysis and image processing. The connected time contour was used in order to provide users as much information as possible on the nature of the task.

For all 60 high-speed video sequences, participants were asked to rate the video quality (lighting conditions, contrast relative to the discrimination of the glottal area), as well as the segmentation quality (tracking of vocal-fold movements, irrelevant excrescences). The user was presented with a sequence of frames which were consecutively evaluated. If the contour correctly followed the glottal area, the user advanced to the following frame. In the contrary case, the user could control the contour and correct it using the mouse. Two examples of the interactive tool are presented in Figs. 8 and 9. When the sequence processing terminated, the user was asked to evaluate the video and segmentation quality on a 5-point scale, with 1 representing very bad quality and 5 very good quality. On average, each sequence required about 15 min to be fully processed.

4.3. Automatic threshold detection

The proposed automatic glottal detection method has been compared to a fully-automatic threshold-based method. We selected the most recent one, presented by Mehta et al. (2010), Mehta et al. (2011). Similarly to the method used by Mehta et al. (2010), Mehta et al. (2011), the intensity threshold for each glottal cycle is estimated as the minimum between the first two peaks of a smoothed intensity histogram. The high-speed sequence processing is kept identical up to the segmentation procedure. However, for the purpose of comparison, no post-processing user adjustment of the threshold was made.

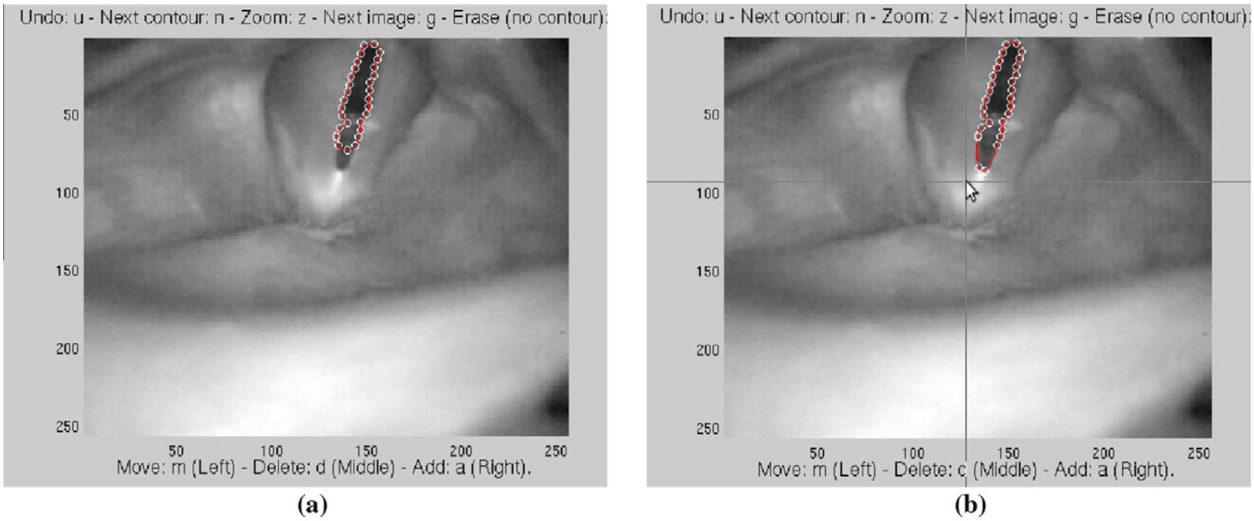

Undo: u - Next contour: n - Zoom: z - Next image: g - Erase (no contour):
50
100
150
200
250
50
100
150
200
250
Move: m (Left) - Delete: d (Middle) - Add: a (Right).
(a)
Undo: u - Next contour: n - Zoom: z - Next image: g - Erase (no contour):
50
100
150
200
250
50
100
150
200
250
Move: m (Left) - Delete: d (Middle) - Add: a (Right).
(b)

Fig. 8. Segmentation errors and manual evaluation of early convergence: (a) computed contour on a image; the anterior part of the glottis has not been detected and (b) corrected contour; the contour now tracks the entire glottal area.

Please cite this article in press as: Karakozoglou, S.Z. et al., Automatic glottal segmentation using local-based active contours and application to glottovibrography, Speech Comm. (2011), doi:10.1016/j.specom.2011.07.010

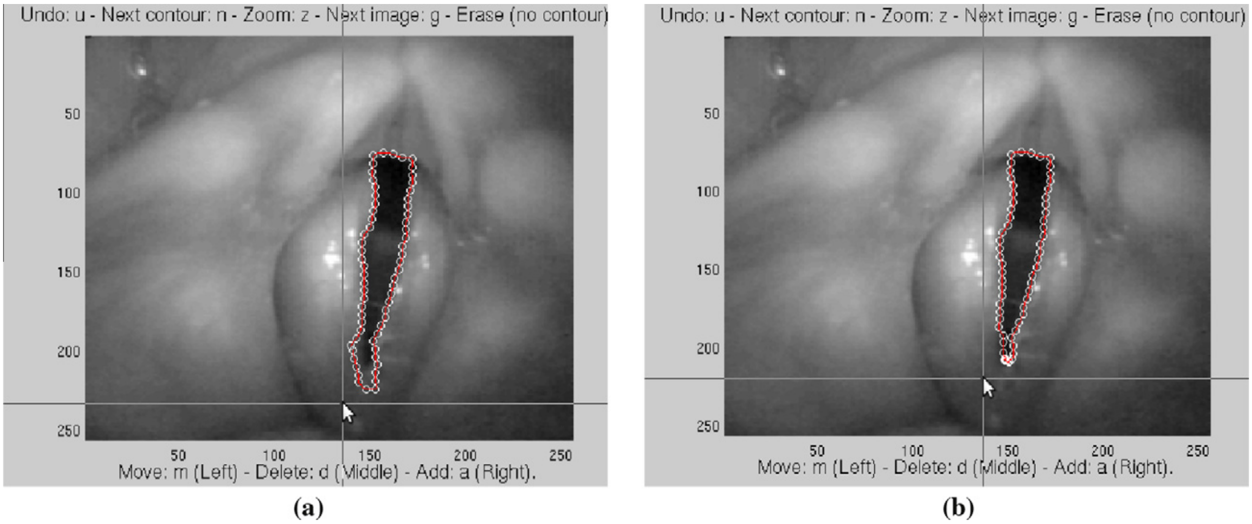

Undo: u - Next contour: n - Zoom: z - Next image: g - Erase (no contour)
50
100
150
200
250
50
100
150
200
250
Move: m (Left) - Delete: d (Middle) - Add: a (Right).
(a)
Undo: u - Next contour: n - Zoom: z - Next image: g - Erase (no contour)
50
100
150
200
250
50
100
150
200
250
Move: m (Left) - Delete: d (Middle) - Add: a (Right).
(b)

Fig. 9. Segmentation errors and manual evaluation of excrescences: (a) computed contour on a image; the contour includes an irrelevant anterior region and (b) corrected contour; the contour now tracks the entire glottal area.

5. Results

5.1. Evaluation

5.1.1. Comparison between manual and automatic segmentation

The manual verification of the automatic segmentation resulted in a number of interesting findings. Concerning the subjective evaluation on a 5-point scale (1 – very bad quality to 5 – very good quality), video quality was rated 4.2 ± 0.7 (mean value ± standard deviation) and segmentation quality 4.2 ± 1.0. In 71% of the sequences, the segmentation was rated equal to or higher than the video quality, while 93% were characterized as more than acceptable (average to very good).

Concerning the qualitative evaluation based on manual correction of the segmentation, the absolute segmentation error (by considering the absolute differences) for the entire database was 4 ± 18 pixels (mean value ± standard deviation).

The majority of segmentation errors occurred either in the posterior or anterior part of the glottal area. Either the active contours converged before the actual vocal-fold edges, or they included surrounding areas. As evaluated in most cases, the contour satisfactorily tracked the glottal area, despite errors around the edges. Two cases where the segmentation procedure failed to effectively track the glottal area are presented in Figs. 8 and 9.

An important aspect in evaluating the segmentation results concerns the static phases of glottal opening and closing instants. The segmentation procedure needs to meet the demands for the glottal source analysis. For that reason, the amount of error relative to the glottal instants was also investigated. The absolute segmentation error on closing and

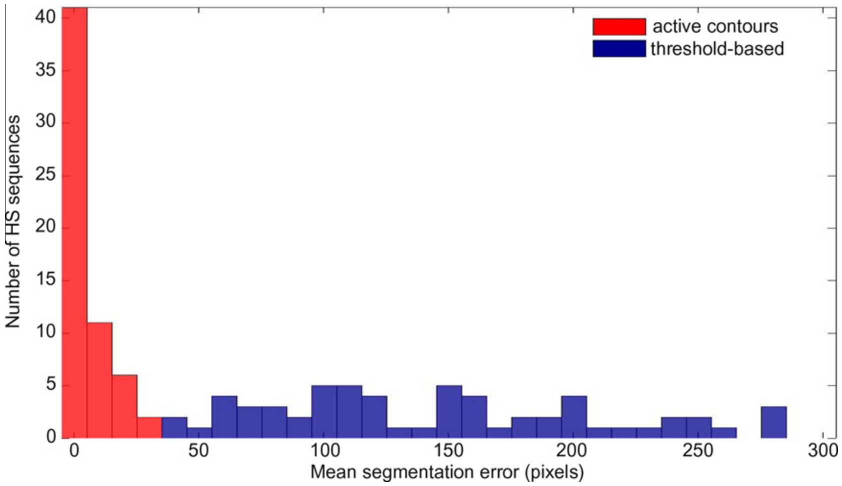


Fig. 10. Glottal detection error from the active contour method (red histogram) and the threshold-based one (blue histogram). Active contour method presents significantly lower error of segmentation. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Please cite this article in press as: Karakozoglou, S.Z. et al., Automatic glottal segmentation using local-based active contours and application to glottovibrography, Speech Comm. (2011), doi:10.1016/j.specom.2011.07.010

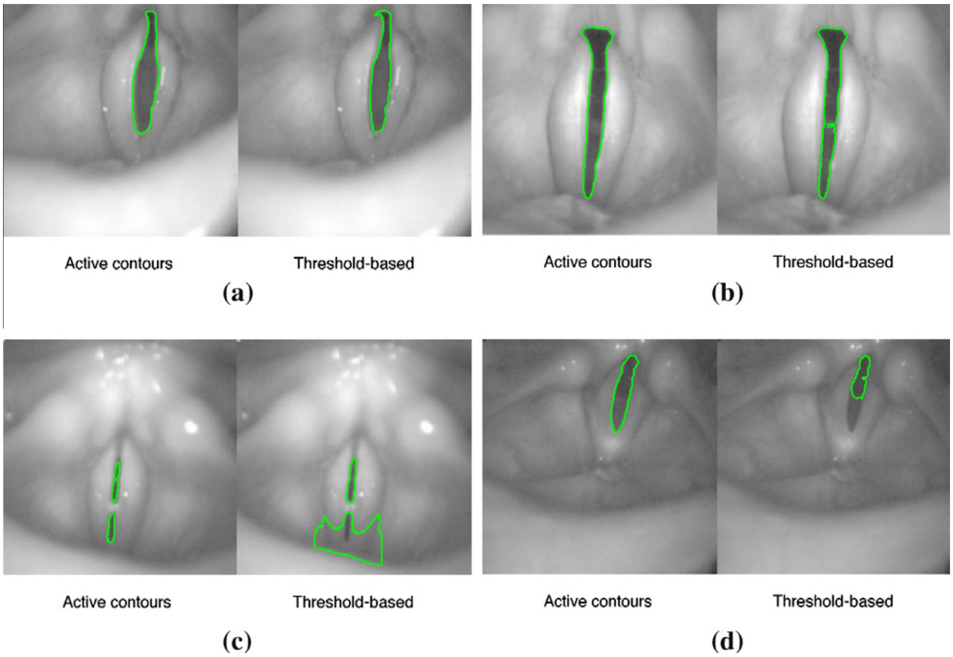


Fig. 11. Four examples of comparisons of glottal detection between active contour (left) and threshold-based method (right).

opening instants on the entire database was 3 ± 10 pixels (mean value ± standard deviation), based on results acquired from the manual correction of the segmentation.

5.1.2. Comparison with threshold-based detection

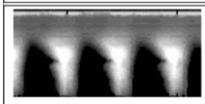
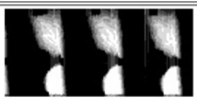
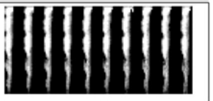
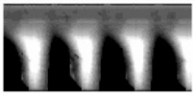
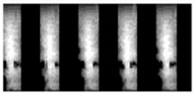
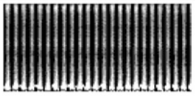
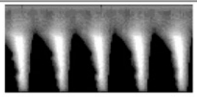
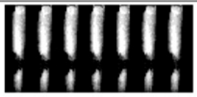
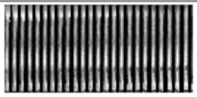
The active contour method was evaluated in comparison to fully-automatic threshold-based glottal detection. Fig. 10 shows histograms of the mean segmentation errors for both methods. It is clear that the active contour method has very low error. Indeed, the relative error in glottal detection between the active contour method and the manual one was 1.35 ± 5%, whereas the relative error between the threshold-based method and the manual one is 69 ± 63%. As already mentioned by Mehta et al. (2011), the threshold-based method usually requires user intervention by visual validation or manual threshold adjustment in order to deal with illumination inconsistency and arytenoid hooding, characterized by the standard deviation of this method's error. When applied in an fully automatic way, without considering the continuity of the vibratory pattern, this method results in false or imprecise area detections. Fig. 11 presents four examples of glottal detection comparisons between active contour and threshold-based method on frames with maximal glottal opening. In Figs. 11(a),(b) results are comparable. Fig. 11(c) presents a case where threshold-based detection includes a large inconsistent surrounding area and the latter shows a case where threshold-based detection does not detect the anterior part of the glottis.

5.2. Application to glottovibrography

The GVG representation is a compact form of assessment of the entire database. Table 1 depicts characteristic GVG representations derived from the database. Each column represents different samples from the same glottal behavior. GVG thumbnails for the entire database are included in the Appendix.

Using this representation, deflections of the glottal area are depicted within a single image. By computing the distance of the vocal-fold edges, the exact physiological behavior of the vocal folds is clearly evident. From the posterior to the anterior part of the glottis, we can observe the shape of the vocal-fold edges and visualize on a single image the glottal dynamic of an entire HSV sequence. The GVG visualization can delineate the vocal-folds

Table 1
Characteristic glottal behaviors observed in the HSV database (first row) and corresponding GVG images. These prototypic examples are extracted from the database presented in Table A.1.

Partial Glottal Closing, Posterior Glottal Chink	Full Glottal Closing, No/Limited Medial Opening	Full Glottal Closing, High Frequency
		
		
		

vibratory pattern in cases where the PVG fails to be clear due to poor detection of the glottal axis. Erroneous placement of the glottal axis results in miscalculation of the deflection of vocal-fold edges from the glottal axis. Therefore, the PVG representation is blurry, while the GVG captures more precisely the deflection of vocal-fold edges. By implementing GVG and PVG algorithms for segmented HSV sequences, we can observe the representation differences of the two visualizations. In Figs. 12 and 13 two examples are shown, where the vibratory pattern is more

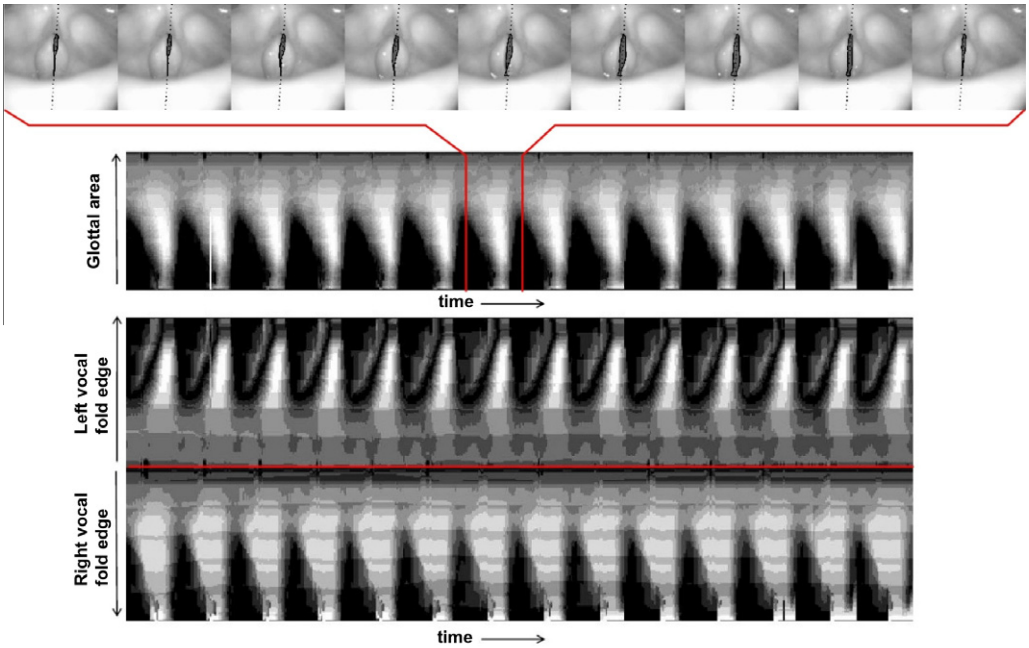


Fig. 12. GVG (upper panel) and PVG (lower panel) in a case of partial glottal closing. The opening and closing phases are distinguishable. The red horizontal line in the PVG represents the posterior part of the vocal-fold edges. Within one glottal cycle, marked between the red lines in the GVG, full frames of the sequence along with the glottal contour and the glottal axis are presented. One out of every four frames are shown for clarity. Sequence 14; see Appendix. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

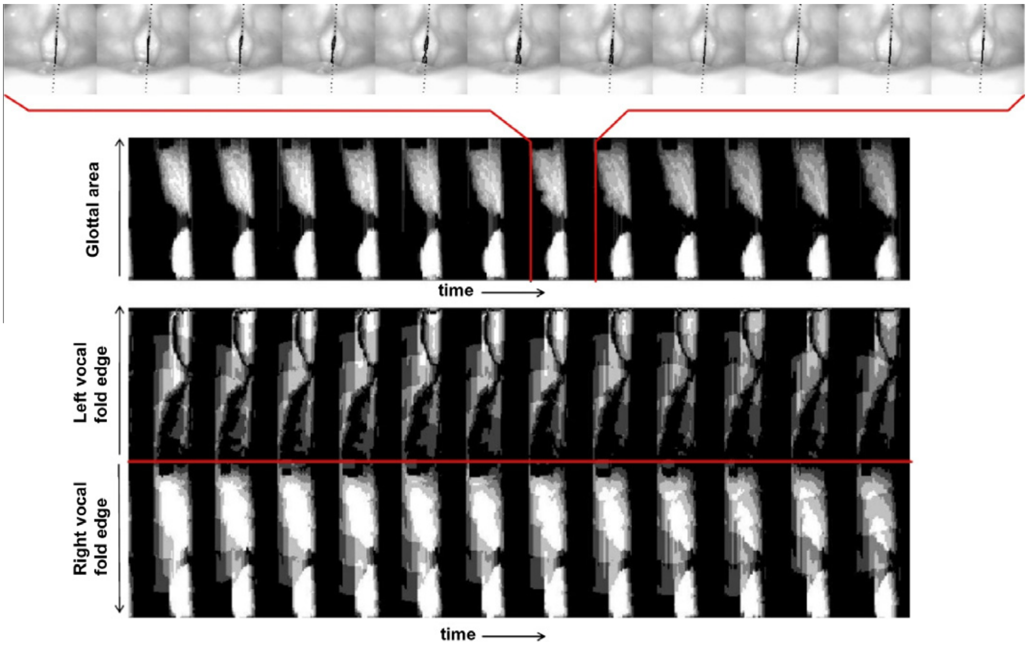


Fig. 13. GVG (upper panel) and PVG (lower panel) in a case of no medial opening. The red horizontal line in the PVG represents the posterior part of the vocal-fold edges. Within one glottal cycle, marked between the red lines in the GVG, full frames of the sequence along with the glottal contour (continuous black line) and the glottal axis (dotted black line) are presented. One out of every four frames are shown for clarity. Sequence 16; see Appendix. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Please cite this article in press as: Karakozoglou, S.Z. et al., Automatic glottal segmentation using local-based active contours and application to glottovibrography, Speech Comm. (2011), doi:10.1016/j.specom.2011.07.010

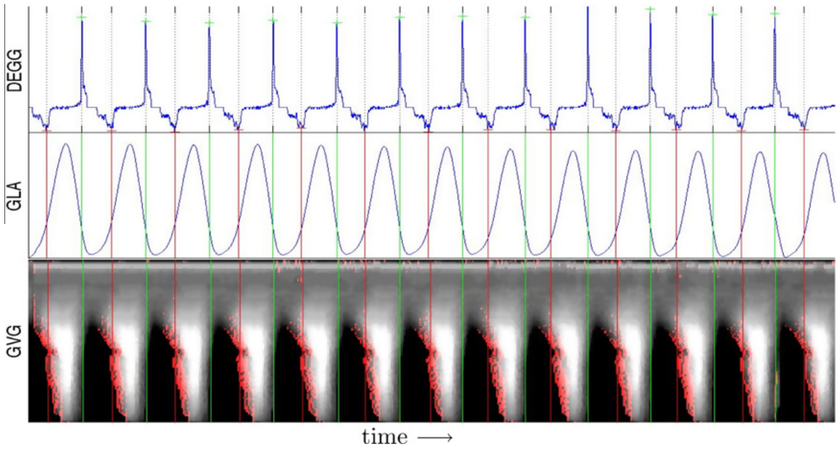


Fig. 14. Synchronous representation of DEGG, GLA signals and GVG with maximum speed profile in the case of an incomplete glottal closing. Sequence 2; see Appendix.

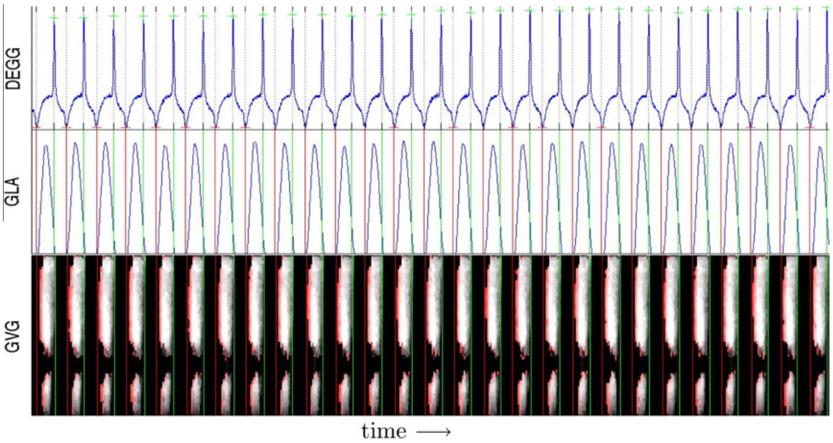


Fig. 15. Synchronous representation of DEGG, GLA signals and GVG with maximum speed profile in the case of no medial opening. Sequence 25; see Appendix.

distinctly represented in the GVG rather than the corresponding PVG representation. Fig. 12 presents a case of partial glottal closing. Fig. 13 presents a case of no medial opening. This is a case of a vibratory pattern which consists of the zipper-like movement in two distinct regions. In the PVG representation, the left and right vocal-fold edges are not distinct due to the false placement of the glottal axis. In contrast, in the GVG representation, the two regions and their evolution are easier to observe. The asymmetry of the vibration is evident in the GVG representation; the vocal folds close faster than they open. The two glottal regions never merge completely.

6. Conclusion and perspectives

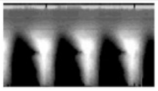
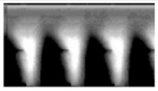
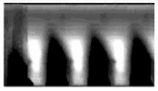
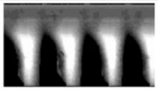
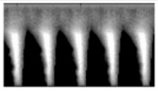
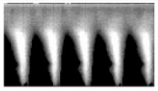
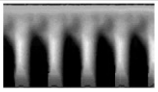
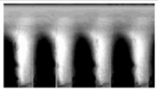
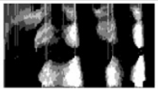
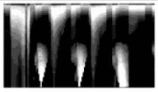
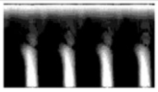
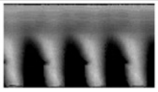
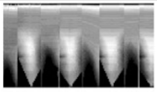
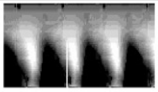
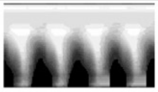
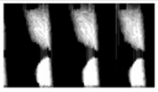
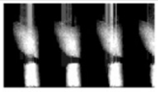
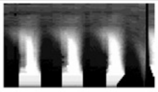
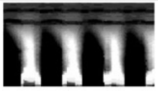
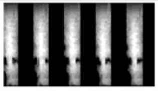
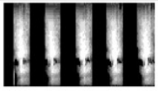
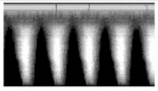
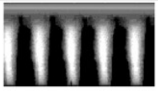
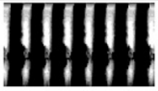
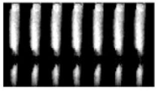
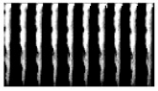
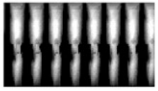
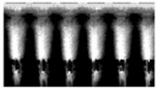
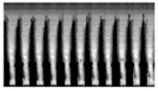
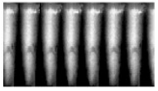
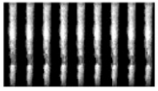
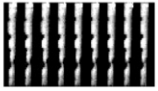
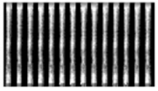
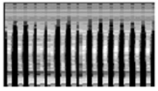
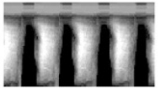
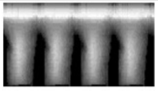
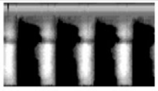
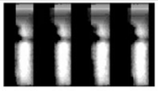
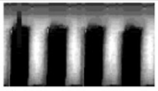
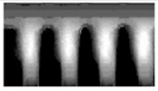
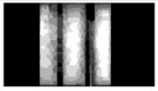
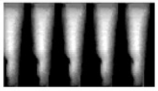
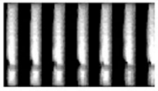
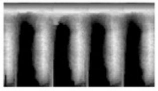
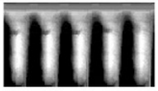
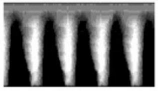
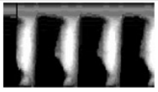
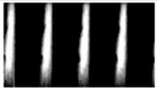
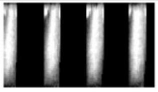
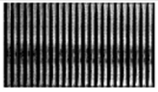
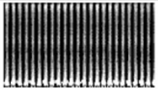
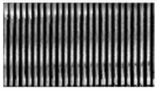
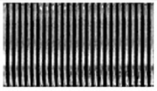
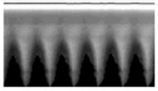
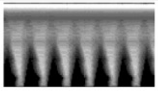
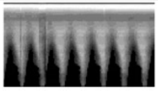
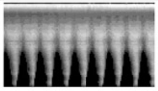
The most powerful way of precisely capturing vocal-fold vibrations is high-speed videoendoscopy. We present a new segmentation method which consists of a local-based active contour method for automatic segmentation with no user intervention. The local-based scheme is the first to be used in glottis segmentation, taking into consideration the inhomogeneous nature of medical images. The presented framework consists of distinct steps, which are optimized to give maximum information. A precise scheme for glottis localization is introduced, which facilitates the automatic segmentation procedure. The local-based framework deals effectively with various image qualities and surrounding clutter and is versatile enough to split and merge, so as to track the various vibratory patterns.

The use of the presented glottovibrogram proposes an effective and compact alternative to the time-consuming visualization of high-speed sequences. GVG, being an improvement of previous visualizations, addresses the problem of glottal-axis detection by proposing a representative form of vocal-fold vibration.

A database of 60 recordings of high-speed sequences is used to investigate the segmentation's discriminative power, that is its ability to correctly discriminate the object of interest from the background. Validation of the glottis segmentation scheme is made by manual verification and by automatic threshold-based glottal detection. GVG is compared to the existing PVG representation and is shown

Please cite this article in press as: Karakozoglou, S.Z. et al., Automatic glottal segmentation using local-based active contours and application to glottovibrography, Speech Comm. (2011), doi:10.1016/j.specom.2011.07.010

Table A.1
GVG thumbnails of the UKE database, corresponding to 35 ms. The high-speed sequence indices are shown below each thumbnail with information on the speech or singing condition.

 1. normal	 2. normal	 3. normal	 4. normal	 5. normal	 6. normal
 7. breathy	 8. breathy	 9. creaky	 10. pressed	 11. pressed	 12. breathy
 13. breathy	 14. breathy	 15. breathy	 16. creaky	 17. creaky	 18. pressed
 19. pressed	 20. pitch D3 in M1	 21. pitch D3 in M1	 22. pitch D3 in M1	 23. pitch D3 in M1	 24. pitch A3 in M1
 25. pitch A3 in M1	 26. pitch D4 in M1	 27. pitch D4 in M1	 28. glissando	 29. glissando	 30. glissando
 31. pitch A3 in M2	 32. pitch D4 in M2	 33. pitch D4 in M1	 34. pitch A4 in M2	 35. pitch A4 in M2	 36.breathy
 37. breathy	 38. normal	 39. normal	 40. breathy	 41. breathy	 42. creaky
 43. creaky	 44. glide down	 45. glide down	 46. breathy	 47. breathy	 48. glide up
 49. glide up	 50. pressed	 51. pressed	 52. pressed	 53. pitch F4# in M2 soft vibrato	 54. pitch F4# in M2 soft vibrato
 55.pitch F4# in M2 strong vibrato	 56. pitch F4# in M2 soft vibrato	 57. breathy glide down	 58. breathy glide down	 59. breathy glide down	 60. normal glide up

to provide a clear representation of vocal-fold vibration while reducing error from glottal-axis detection.

Several data may be extracted from high-speed sequences, such as GVG or the relative glottal area (GLA) computed on high-speed sequences by the sum of pixels assigned to the detected glottal area in each frame. Results can be compared to other glottal-activity signals for a more complete assessment of vocal-fold vibration, for evaluating the use of high-speed data for voice processing and the nature of information. EGG, for instance, is the most common non-invasive investigation technique of glottal contact. Its derivative (DEGG) can advantageously be used for the analysis of glottal activity (Henrich et al., 2004).

The recordings used in this study also included EGG signals, which enable us to simultaneously represent DEGG, GLA and GVG signals. Figs. 14 and 15 present a comparison between the glottal closing and opening instants measured on DEGG signals and the features observed on the detected glottal area (GLA signal), as well as on the GVG visualization with maximum speed profile. The green crosses and lines indicate the positions of glottal closing instants (GCI), while the red marks indicate the positions of glottal opening instants (GOI). Glottal closing instants are defined as the instants the glottal area decreases with highest velocity (Childers et al., 1990; Childers, 1995). They are also related to abrupt increases in glottal contact and coincide with a local maximum in the DEGG signal (Henrich et al., 2004). In both cases here, which present abrupt closures, a strong coincidence is observed between the DEGG peaks and the rapid increase in glottal contact. Similarly, when glottal opening is abrupt, as is the case in Fig. 15, GOIs are related to the instant of decrease in contact. However, when glottal opening is smoother, as in Fig. 14 where gradual posterior-to-anterior openings can be observed, GOIs are related to the instants of decrease in contact in the median part of the glottis. These images represent a first attempt towards bridging EGG and image based glottal analysis. Future work will be devoted to systematic comparisons of EGG and GVG representations.

Acknowledgements

The authors would like to thank Frank Müller, Götz Schade and Markus Hess for the database collection, as well as Brian FG Katz for proof reading the article.

Appendix A

The high-speed database used in this paper is presented in Table A.1, by means of the new GVG visualization of glottal dynamics introduced in Section 3. Time is represented on the *x*-axis. Glottal-edges distance along the vocal-fold length is represented on the *y*-axis, from the anterior part (bottom of each image) to the posterior part (top of each image) of the glottis. Black corresponds to zero distance and white to maximal distance. Subject S1 produced speech samples only (sequences 1 to 19 and 57 to 60). Subject S2 produced singing samples (sequences 20 to 35, 53 to 56) and speech samples (sequences 36 to 52). Speech samples consisted of sustaining a given voice quality (breathy, normal, pressed, creaky), and producing up and down glides. Singing samples consisted of glissandos and sustained sounds at a given pitch and laryngeal mechanism (M1 is synonymous of modal laryngeal register, M2 of falsetto one. See Roubeau et al., 2009).

References

Adams, R., Bischof, L., 1994. Seeded region growing. IEEE Trans. Pattern Anal. Mach. Intell. 16, 641–647.

Allin, S., Galeotti, J., Stetten, G., Dailey, S., 2004. Enhanced snake based segmentation of vocal folds. In: Proc. IEEE Int. Symp. Biomed. Imaging, pp. 812–815.

Bailly, L., 2009. Interaction entre cordes vocales et bandes ventriculaires en phonation: exploration in-vivo, modélisation physique, validation in-vitro. Ph.D. thesis. Université du Maine.

Bezier, P., 1972. Numerical Control; Mathematics and Applications. John Wiley & Sons.

Chan, T., Vese, L., 2001. Active contours without edges. IEEE Trans. Image Process. 10, 266–277.

Childers, D., 1995. Glottal source modeling for voice conversion. Speech Commun. 16, 127–138.

Childers, D., Hicks, D., Moore, G., Eskenazi, L., Lalwani, A., 1990. Electroglottography and vocal fold physiology. J. Speech Lang. Hear. Res 33, 245.

Childers, D., Krishnamurthy, A., 1985. A critical review of electroglottography. Crit. Rev. Biomed. Eng. 12, 131–161.

Deliyski, D., Petrushev, P., 2003. Methods for objective assessment of high-speed videoendoscopy. In: Proc. 6th Int. Conf. Adv. in Quant. Laryngol. Voice Speech. Res AQL, pp. 1–16.

Deliyski, D., Petrushev, P., Bonilha, H., Gerlach, T., Martin-Harris, B., Hillman, R., 2008. Clinical implementation of laryngeal high-speed videoendoscopy: challenges and evolution. Folia Phoniatr. Logop. 60, 33–44.

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

Döllinger, M., Lohscheller, J., Svec, J., McWhorter, A., Kunduk, M., 2011. Advances in Vibration Analysis Research. Farzad Ebrahimi. chapter Support Vector Machine Classification of Vocal Fold Vibrations Based on Phonovibrogram Features. pp. 435–456.

Einig, D., 2010. Merkmalsbasierte Beschreibung von Phonovibrogrammen bei pathologischer Stimmgebung durch Entwicklung eines Frameworks zur Analyse medizinischer Datenmodalitäten (Feature-based Description of Phonovibrograms in Pathological Phonation through the Development of a Framework for the Analysis of Medical Data Modalities). Master's thesis. Trier University of Applied Sciences.

Gilbert, H., Potter, C., Hoodin, R., 1984. Laryngograph as a measure of vocal fold contact area. J. Speech Hear. Res. 27, 178–182.

Glasbey, C., 1993. An analysis of histogram-based thresholding algorithms. Graph. Models 55, 532–537.

Haralick, R., Shapiro, L., 1985. Image segmentation techniques. Comput. Graph. Image Process. 29, 100–132.

Henrich, N., d'Alessandro, C., Doval, B., Castellengo, M., 2004. On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation. J. Acoust. Soc. Am. 115, 1321–1332.

Henrich, N., 2001. Etude de la source glottique en voix parlée et chantée: modélisation et estimation, mesures acoustiques et électroglottographiques, perception (Study of the glottal source in speech and singing: Modeling and estimation, acoustic and electroglottographic measurements, perception). Ph.D. thesis. Université Pierre et Marie Curie - Paris 6.

Karakozoglou, S.Z., 2010. Glottal source analysis: a combinatory study using high-speed videoendoscopy and electroglottography. Master's thesis. Université Paris-Sud XI, University of Crete.

Kass, M., Witkin, A., Terzopoulos, D., 1988. Snakes: Active contour models. Int. J. Comput. Vis. 1, 321–331.

Kohler, R., 1981. A segmentation system based on thresholding. Comput. Graph. Image Process. 15, 319–338.

Lankton, S., Tannenbaum, A., 2008. Localizing region-based active contours. IEEE Trans. Image Process. 17, 2029–2039.

Lohscheller, J., Döllinger, M., Schuster, M., Schwarz, R., Eysholdt, U., Hoppe, U., 2004. Quantitative investigation of the vibration pattern of the substitute voice generator. IEEE Trans. Biomed. Eng. 51, 1394–1400.

Lohscheller, J., Eysholdt, U., Toy, H., Döllinger, M., 2008. Phonovibrography: Mapping high-speed movies of vocal fold vibrations into 2-D diagrams for visualizing and analyzing the underlying laryngeal dynamics. IEEE Med. Imag. Trans. 27, 300–309.

Lohscheller, J., Toy, H., Rosanowski, F., Eysholdt, U., Döllinger, M., 2007. Clinically evaluated procedure for the reconstruction of vocal fold vibrations from endoscopic digital high-speed videos. Med. Image Anal. 11, 400–413.

Marendic, B., Galatsanos, N., Bless, D., 2001. A new active contour algorithm for tracking vibrating vocal folds. In: Proc. Int. Conf. Image Proc., pp. 397–400.

Mehnert, A., Jackway, P., 1997. An improved seeded region growing algorithm. Pattern Recognit. Lett. 18, 1065–1071.

Mehta, D., Deliyski, D., Quatieri, T., Hillman, R., 2011. Automated measurement of vocal fold vibratory asymmetry from high-speed videoendoscopy recordings. J. Speech Lang. Hear Res 54, 47–54.

Mehta, D., Deliyski, D., Zeitels, S., Quatieri, T., Hillman, R., 2010. Voice production mechanisms following phonosurgical treatment of early glottic cancer. Ann. Otol. Rhinol. Laryngol. 119, 1–9.

Moukalled, H., Deliyski, D., Schwarz, R., Wang, S., 2009. Segmentation of laryngeal High-Speed Videoendoscopy in temporal domain using paired active contours. MAVEBA 1, 137–140.

Neubauer, J., Mergell, P., Eysholdt, U., Herzel, H., 2001. Spatio-temporal analysis of irregular vocal fold oscillations: Biphonation due to desynchronization of spatial modes. J. Acoust. Soc. Am. 110, 3179–3192.

Rothenberg, M., 1992. A multichannel electroglottograph. J. Voice 6, 36–43.

Roubeau, B., Henrich, N., Castellengo, M., 2009. Laryngeal vibratory mechanisms: The notion of vocal register revisited. J. Voice 23, 425–438.

Samet, H., Tamminen, M., 1988. Efficient component labeling of images of arbitrary dimension represented by linear bintrees. IEEE Trans. Pattern Anal. Mach. Intell. 10, 586.

Scherer, R., Druker, D., Titze, I., 1988. Electroglottography and direct measurement of vocal fold contact area. In: O, F. (Ed.), Vocal fold physiology: voice production, mechanisms and functions. New York: Raven, pp. 279–291.

Sethian, J., 1999. Level Set Methods and Fast Marching Methods: Evolving Interfaces in Computational Geometry, Fluid Mechanics, Computer Vision, and Materials Science. Cambridge University Press.

Westphal, L., Childers, D., 1983. Representation of glottal shape data for signal processing. IEEE Trans. Acoust. 31, 766–769.

Yan, Y., Chen, X., Bless, D., 2006. Automatic tracing of vocal-fold motion from high-speed digital images. IEEE Trans. Biomed. Eng., 53.

Zuiderveld, K., 1994. Contrast limited adaptive histogram equalization. In: Graphics gems IV, pp. 474–485.

Please cite this article in press as: Karakozoglou, S.Z. et al., Automatic glottal segmentation using local-based active contours and application to glottovibrography, Speech Comm. (2011), doi:10.1016/j.specom.2011.07.010