Hindawi Publishing Corporation
Journal of Applied Mathematics
Volume 2014, Article ID 818415, 16 pages
http://dx.doi.org/10.1155/2014/818415

Research Article

Automatic Segmentation of High Speed Video Images of Vocal Folds

Turgay Koç[1,2] and Tolga Çiloğlu[2]

[1] *Department of Electronic Communication Engineering, Süleyman Demirel University, 03200 Isparta, Turkey*
[2] *Department of Electrical and Electronics Engineering, Middle East Technical University, 06800 Ankara, Turkey*

Correspondence should be addressed to Turgay Koç; turgaykoc@sdu.edu.tr

Received 24 January 2014; Revised 18 April 2014; Accepted 20 April 2014; Published 5 June 2014

Academic Editor: Feng Gao

An automatic method for segmenting glottis in high speed endoscopic video (HSV) images of vocal folds is proposed. The method is based on image histogram modeling. Three fundamental problems in automatic histogram based processing of HSV images, which are automatic localization of vocal folds, deformation of the intensity distribution by nonuniform illumination, and ambiguous segmentation when glottal gap is small, are addressed. The problems are solved by using novel masking, illumination, and reflectance modeling methods. The overall algorithm has three stages: masking, illumination modeling, and segmentation. Firstly, a mask is determined based on total variation norm for the region of interest in HSV images. Secondly, a planar illumination model is estimated from consecutive HSV images and reflectance image is obtained. Reflectance images of the masked HSV are used to form a vertical slice image whose reflectance distribution is modeled by a Gaussian mixture model (GMM). Finally, estimated GMM is used to isolate the glottis from the background. Results show that proposed method provides about 94% improvements with respect to manually segmented data in contrast to conventional method which uses Rayleigh intensity distribution in extracting the glottal areas.

1. Introduction

Vocal system monitoring is essential for clinical analysis of voicing and investigation of speech production models. Today, high speed endoscopic video (HSV) of vocal folds is a state-of-the-art method to investigate vocal fold vibration. With the development of high speed video cameras, the vibration of vocal folds can be captured at 4000 fps with an image resolution of 256×256 pixels. Such a system can provide 20 frames in a glottal cycle of 200 Hz vibration. Reduction of the dimensionality of spatiotemporal information and representation of high speed video data in a simple, convenient, and lossless manner is a challenge. By means of HSV, some of the characteristics of vocal fold vibrations, such as asymmetric vibration, glottal area, and glottal width, are measured for clinical and engineering applications. Quantification of the difference between vibration patterns of left and right vocal folds in both spatial and temporal domains is required for objective analysis of voice disorders [1–4]. The time variation of glottal area is of particular interest in HSV based analysis [2, 5–9]. It is used as a reference signal in estimating the parameters of biomechanical models of vocal folds [5, 6]. These models are used for functional analysis of vocal folds vibrations as well as articulatory speech synthesis.

Currently, glottal area extraction from HSV images is based on histogram thresholding [1, 10], region-growing [7, 11], and active contours [12]. Histogram thresholding method uses the difference between intensity distributions of the object and background. In HSV images, pixels corresponding to vocal folds and tissue around them have large intensity values compared to the pixels corresponding to the opening between the vocal folds, that is, glottis. The aim in the histogram based methods is to find a threshold to discriminate the low intensity pixels corresponding to glottis from the high intensity pixels. The intensity distributions are usually modeled by parametric functions. First, distributions of the glottis and background intensities are estimated and a threshold is determined. Glottal area can be estimated quite accurately provided that the intensity distribution is bimodal.

However, intensity distribution is very sensitive to illumination. Nonuniform illumination can ruin the modality of intensity distributions. Thus, it is a significant problem in the histogram based segmentation methods. The advantages of the method are its suitability for real-time applications due to little computational requirement, and under uniform illumination it is very effective for glottal area extraction.

Region-growing based methods use the histogram thresholding as a first step of the algorithm. After applying thresholding, a binary image is obtained for further processing. In the next step, one of the connected regions is used as an initial seed; then the seed points are propagated up to a state at which the difference in the intensity of the boundary pixels and intensities of their neighbours reaches a certain limit. Region-growing methods require an intelligent algorithm to select multiple seeds in video frames, since vocal folds can have partial contacts during glottal opening and closing phases. Therefore, their performances depend on the combination of histogram thresholding and accurate selection of seed points.

In the active contour based methods, an initial contour for glottis is found by using edge detection operation; then it is shrunk and expanded iteratively to minimize an energy function. One disadvantage is that the convergence requires significant amount of time due to computational burden. Another disadvantage is that the final boundary is affected by the selection of initial contour and noise in the HSV images [8, 12]. It is not suitable for applications requiring processing large number of video frames in a small fraction of time or real-time, such as clinical evaluation. However, segmentation of HSV in an automatic manner is indispensable for applications requiring analysis of large number of frames, such as vocal fold vibration functional analysis and speech production analysis.

Currently, the major challenge is automatic extraction of glottal area [7, 8, 12]. Most of the existing image processing algorithms in the literature are not completely automatic and require user intervention. In [7], the intensity distribution of an HSV image is modeled by Rayleigh distribution and a binary image is obtained by using histogram thresholding according to Bayes's decision rule. After selection of single seed, a region-growing operation is applied to find the final boundaries. However, as it is shown later in this study, modeling the intensity distribution of HSV without considering the region of interest is unreliable and finding an accurate threshold may involve ambiguities. In addition, the results may get worse due to the nonuniform illumination. The automatic algorithms used in [8, 12] are based on active contour method. However, analysis of a single image can take about several minutes with these methods. To extract glottal area in an automatic and efficient way, a histogram based method is suitable due to its computational efficiency. In this paper, three fundamental problems in automatic histogram based HSV processing, which are automatic localization of vocal folds, deformation of the bimodality of intensity distribution by nonuniform illumination, and ambiguous segmentation when glottal gap is small, are addressed. The problems are solved by a novel approach which involves automatic TV-Norm based masking, illumination, and reflectance modeling.

The intensity histogram depends on the region of interest in the image. To obtain a bimodal intensity distribution for segmentation, region of interest must contain glottis and tissue in its proximity. Automatic calculation of region of interest is the first step of automatic vocal fold segmentation system. It is required not only in histogram based algorithms but also in other automatic HSV segmentation algorithms to reduce the amount of processed data to minimize computational complexity and to avoid the possibility of false region determination. In the literature, to the best of our knowledge, automatic vocal fold localization is performed in [9, 12]. In [12], first, the darkest image is determined as a representative HSV image; then an edge detection operation followed by connected component analysis is applied on it. At the final stage, a rectangular mask corresponding to the largest vertical connected region is used as a region of interest in the image. However, it is quite sensitive to noise due to the fact that even little noise in the image can degrade the edges and connected components in the image. In [9], an image sequence related to a glottal cycle is used to determine the glottis. It is assumed that the lowest intensity value is in the glottis. The row-wise and column-wise intensity minima are calculated from each frame in the image sequence. Then, their averages form vertical profile, V_n, and horizontal profile, H_n, vectors. They expected that V_n and H_n have a minimum value between the locations of vocal folds margins. The two neighboring maxima on each side of the minimum are used to locate the margins. It is our experience that V_n and H_n are highly sensitive to illumination and tissue structures in the image. Usually it is not possible to locate glottis clearly with the use of these profiles. The proposed solution is based on the intensity variation caused by the vibratory motion of vocal folds. The intensities of vocal fold edges and glottis change almost periodically during the vibration of vocal folds. It is realized that the largest intensity variation in HSV is observed at the glottis. In this study, a novel masking algorithm is proposed which uses the total variation norm (TV-norm) of HSV image at the automatic mask determination stage. It can be used as a preprocessing step in many vocal fold segmentation systems.

The second problem which is addressed here is the deformation of the modality of the intensity distribution due to the nonuniform illumination. Nonuniform illumination over a scene can be reduced by modeling reflectance and illumination. Intensity can be considered as the product of reflectance and illumination [13]. One solution is to model the illumination and then recover the reflectance of the object by removing the effect of illumination. The estimated reflectance image can be used in segmentation. In this paper, a planar illumination model is proposed for vocal fold monitoring systems. By means of the proposed model, bimodal HSV reflectance image histograms can be acquired from complex multimodal HSV intensity histograms.

In HSV image segmentation systems, images are processed consecutively; each frame is first analyzed and then segmented. However, the vibration of vocal folds brings

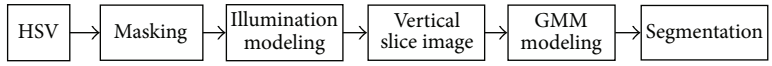


Figure 1: HSV segmentation algorithm.

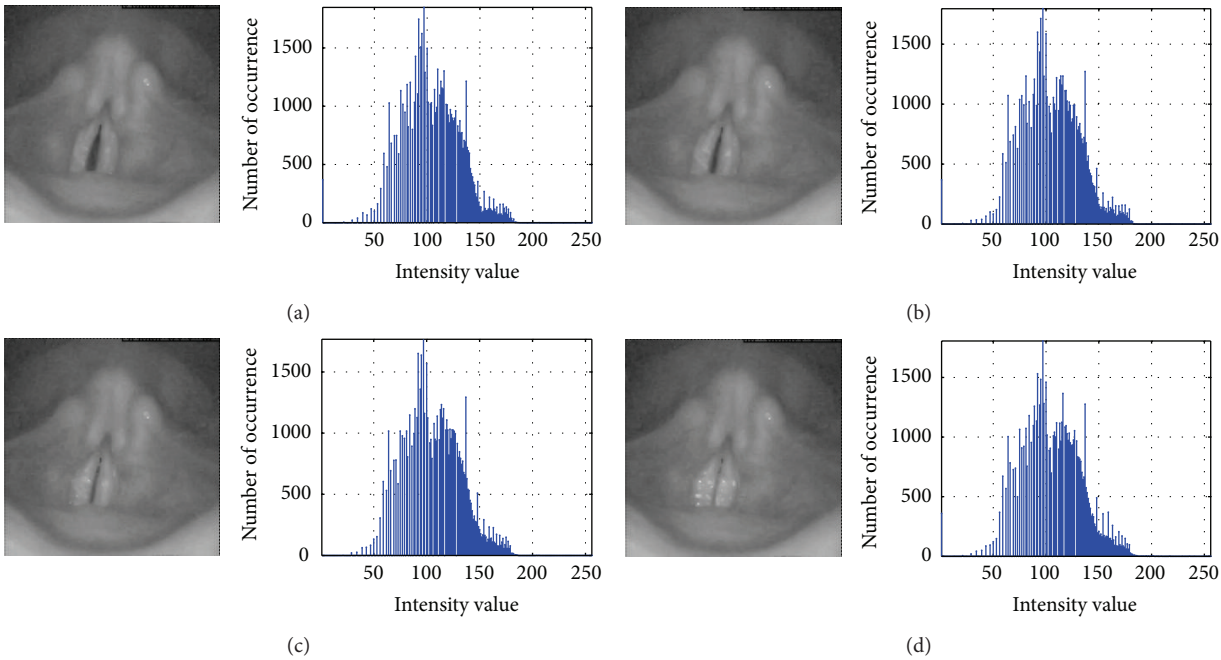


Figure 2: Four snapshots from HSV and corresponding histograms.

some difficulties in histogram based segmentation. The vocal folds' edges approach each other and make a partial or full contact at each glottal period. At the beginning of the glottal opening phase and at the end of the glottal closing phase, a small portion of the image pixels belongs to glottal opening or gap. It causes the intensity histogram of the image to be unimodal which makes finding an accurate and reliable threshold for segmentation difficult. In this study, to remove the uncertainty in threshold determination, a threshold determination algorithm which uses the intensity variation provided by the vibratory motion of vocal folds from an image frame sequence is proposed.

The main steps of the developed algorithm are shown in Figure 1. First, using the TV-norm of an image sequence, a mask is determined to crop a portion of an original frame that contains vocal folds. Then, an illumination model is estimated from the mean temporal intensity variation of the HSV. A reflectance image is estimated after the illumination modeling phase. In the next step, reflectance histogram along central longitudinal cross sections of glottis over a sequence of masked frames is used to determine the threshold for glottal boundary detection. The reflectance distribution is modeled by using Gaussian mixture model (GMM). Finally, in the segmentation step, a reflectance threshold is determined by a Bayesian approach.

This paper is organized as follows. Section 2 describes the automatic vocal fold localization problem and then presents automatic masking algorithm. The degradation of intensity distribution due to the nonuniform illumination and proposed planar illumination model is presented in Section 3. Constructing vertical slice image from masked frames and GMM based reflectance modeling for segmentation are presented in Section 4. Segmentation results and comparison of proposed algorithm with Yan's method [7] on manually marked HSV images are presented in Section 5. Finally, conclusion is given in Section 6.

2. Masking

The accurate determination of region of interest in a HSV image is necessary for intensity histogram processing. Some HSV image frames and corresponding intensity histograms are shown in Figure 2. The vocal folds are located in the middle of the consecutive image frames. The dark region between the folds is called glottis. As the folds come together, the area of the glottis decreases as seen in the images.

Despite the considerable change in the glottal area in each image shown in Figure 2, the intensity histograms are identical and almost unimodal. The intensity distribution is insensitive to the change in the glottal area if the whole HSV image is chosen as a region of interest. By visual inspection of the histograms, it is not easy to decide whether the vocal folds are open or close. Furthermore, due to the almost unimodal intensity distribution, it is hard to determine a reliable threshold value to discriminate the intensity values of pixels of the glottis from the intensity values of its neighbouring

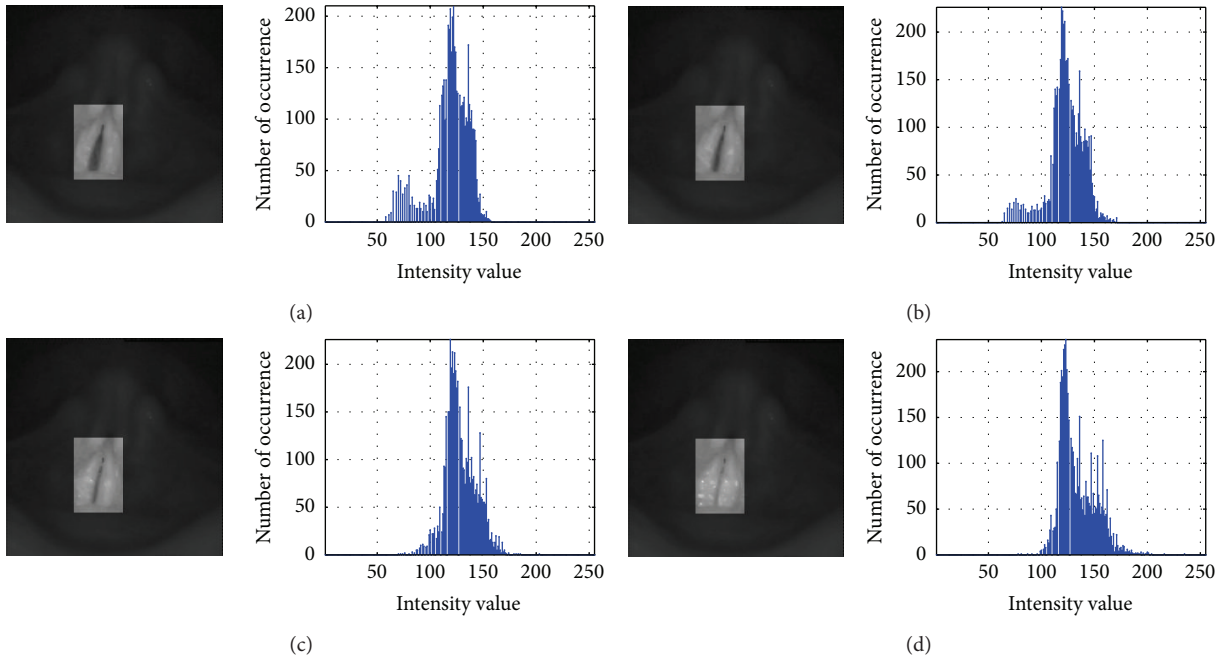


Figure 3: Four snapshots from masked HSV and corresponding histograms.

pixels in the intensity histograms. However, by reducing the region of interest as shown in Figure 3, the intensity histograms can be made bimodal and sensitive to the changes in the glottal area.

The intensity histograms of the masked image frames shown in Figure 3 are either unimodal or bimodal depending on the state of the vocal folds. When the folds are closed or almost closed as shown in Figures 3(c) and 3(d), the small intensity values corresponding to the dark region at the glottis are replaced by the high intensity values of vocal folds edges; hence background intensity distribution is dominant in the corresponding histograms. As a result, the intensity distributions turn out to be unimodal. On the other hand, when the vocal folds are open as shown in Figures 3(a) and 3(b), the small intensity values due to the darkness in the glottis yield the distributions have a stronger bimodal character because of the reduced size of data collection region (by masking). A reliable threshold estimation from an intensity histogram is possible when the opening between the vocal folds is sufficiently large in the masked HSV image. An intensity value at the left edge of the background intensity distribution can be chosen as a threshold [13]. For automatic processing of HSV images, an automatic algorithm that determines the location of the glottis is required. A novel automatic HSV masking algorithm is presented in the following subsection.

2.1. Automatic Masking Algorithm. The algorithm yields a rectangular mask from consecutive HSV images in which vocal folds vibrate. It uses the total variation of the intensity values of each pixel. Total variation norm is used in image restoration and noise reduction [14–16]. In this paper, it is used in the opposite direction. In HSV images, the most active structures are vocal folds. Their vibration produces large intensity changes at pixels corresponding to the glottis (later demonstrated in Section 3). Hence, a large frame to frame intensity variation is a significant indication of a pixel corresponding to glottis.

TV-norm is computed over a sequence of frames as

$$\mathrm{TV}(x,y) = \sum_{n=1}^{N-1} \left| I(x,y,n+1) - I(x,y,n) \right| \quad \forall x, y \in I, \tag{1}$$

where x and y are the spatial variables and n denotes frame index. $I(x, y, n)$ is the intensity function and N is the number of frames in the sequence. It should be emphasized that TV-norm is used in time, not in space. $\mathrm{TV}(x, y)$ produces larger values at the locations of high intensity variation. TV obtained from 300 HSV image frames is shown in Figure 4.

The largest TV values (red color) are accumulated at the glottal region while the smaller values are distributed over the background. This is used for the determination of a mask to locate the glottis in HSV frames automatically.

In placing the mask, horizontal and vertical maxima statistics are used. Let M_x and M_y be two sets whose elements are determined by

$$\begin{aligned} M_x(y) &= \max_x \left(\mathrm{TV}(x,y)\right), \quad y = 1, 2, \ldots, N, \\ M_y(x) &= \max_y \left(\mathrm{TV}(x,y)\right), \quad x = 1, 2, \ldots, N. \end{aligned} \tag{2}$$

M_x and M_y calculated over TV-image in Figure 4 can be seen in Figure 5. The largest M_x values are located between values about 90 and 120. This interval points out the horizontal interval of the most active region, glottis, in the

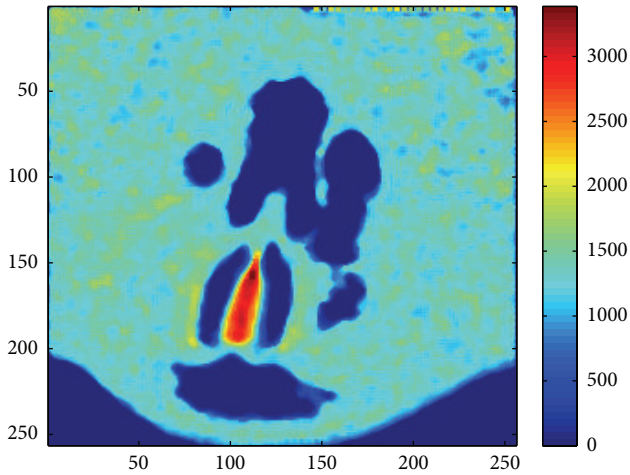


FIGURE 4: TV-image obtained from 300 frames.

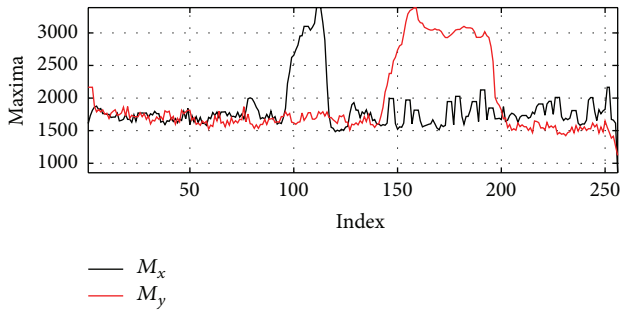


FIGURE 5: Plots of $M_x(y)$ and $M_y(x)$ calculated from TV-image shown in Figure 4.

TV-image. The largest M_y values are between about 150 and 200 and they indicate the minimum and maximum vertical axis points of the location of the glottis in TV image.

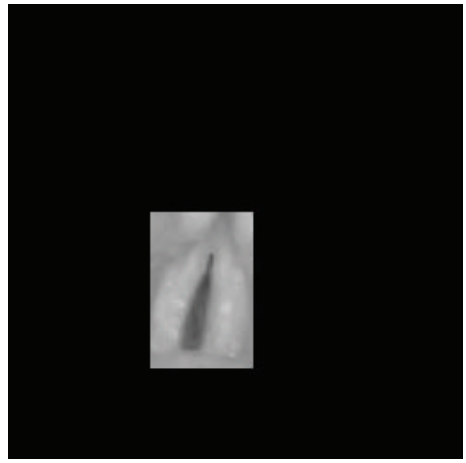
To extract the glottis region from TV-image, a threshold can be determined using M_x and M_y. Let μ and σ be the mean and standard deviation, respectively, of the union of the values of M_x and M_y. TV-image is converted to a binary image by using $T = \mu + \sigma$ as a threshold. Resulting binary images may have nonzero regions, R_i, other than the glottal region, R_g. Figure 6(a) shows the results of the thresholding for the TV-image shown in Figure 4. One large region corresponding to the glottis and several small regions due to noise are seen in the figure. To get rid of these small regions, average TV values inside the regions can be used. Let E_i be the average of the TV values in R_i; that is,

$$E_i = \frac{1}{A_i} \sum_{x,y \in R_i} \mathrm{TV}(x, y), \tag{3}$$

where A_i is the area of the region R_i. By choosing the region having the largest E_i as glottis region the nonzero regions can be eliminated. The index of the glottal region, R_g, is determined as

$$g = \arg\max_i (E_i). \tag{4}$$

(a)



(b)

FIGURE 6: (a) Boundary of the glottal activity region, (b) a masked HSV image.

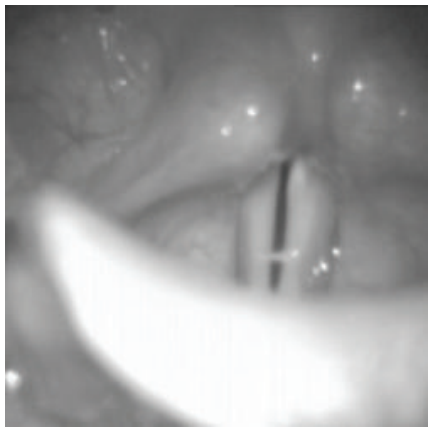
To locate the mask, the boundary, B_g, of the largest TV-norm region, R_g, is determined by an edge detection algorithm. B_g is a set of (x, y) pairs (coordinate values). To construct the mask, the extrema of the elements of B_g are defined as

$$x_{\min} = \arg\min_x (x, y) \in B_g, \qquad y_{\min} = \arg\min_y (x, y) \in B_g,$$
$$x_{\max} = \arg\max_x (x, y) \in B_g, \qquad y_{\max} = \arg\max_y (x, y) \in B_g. \tag{5}$$

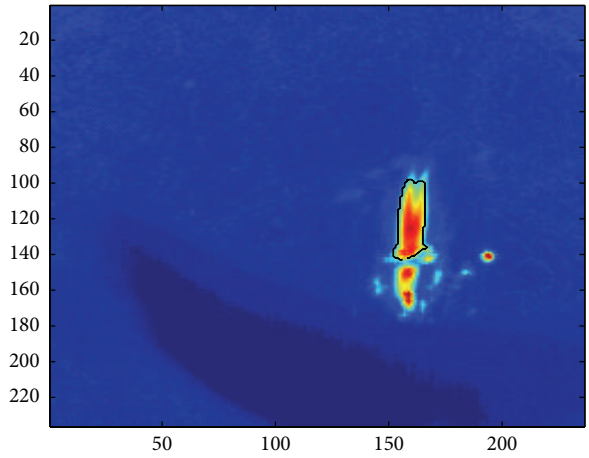
Then the mask is defined as follows:

$$\mathrm{mask}(x, y) = \begin{cases} 1, & x_{\min} \le x \le x_{\max}, y_{\min} \le y \le y_{\max}, \\ 0, & \text{elsewhere.} \end{cases} \tag{6}$$

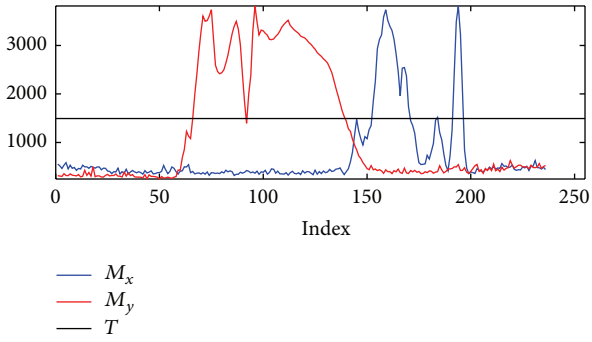
Masked image is obtained by the product of the original image and the mask $I_{\mathrm{MASKED}}(x, y, n) = I(x, y, n) \cdot \mathrm{mask}(x, y)$. An example of a glottal boundary and the corresponding masked image are shown in Figure 6. The vocal folds are cropped from the HSV image; thus the masked image



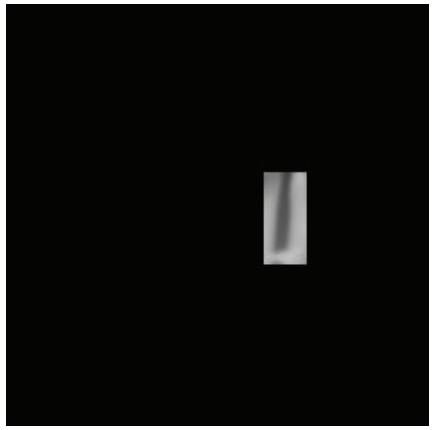
(a) HSV image when glottis is separated into two regions



(b) TV-image and obtained contour



(c) M_x and M_y curves calculated from the TV-images in (b)



(d) Masked HSV

FIGURE 7: Application of the proposed method without 2D median filtering when glottis is separated into two regions.

contains only glottis and its neighbours as desired. This mask is used for processing all 300 frames. The size and location of the mask can be updated for long HSV recordings by processing sliding HSV image blocks containing a number of consecutive frames.

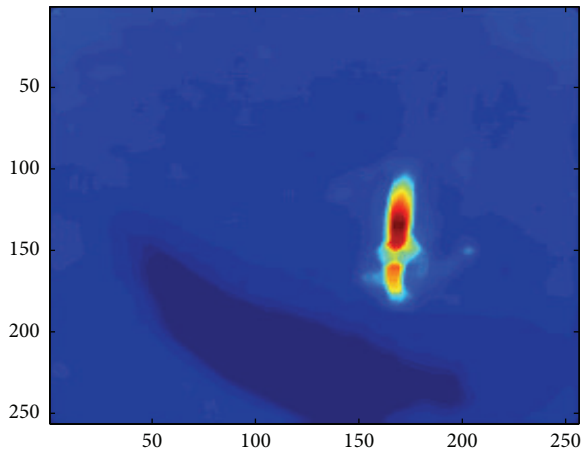
2.1.1. Special Cases

Case 1 (finding ROI when glottis has more than one area). A special case in vocal fold vibration is separation of glottis into more than one vibrating region. An example from the IRCAM database is seen in Figure 7(a). The glottis is separated into two regions. In this special case, the movement of the middle part of vocal folds is limited. The TV-image obtained from Figure 7(a) is shown in Figure 7(b). The total variation is high at both glottal regions but quite small at the connection of the regions. It is also seen in the maxima curves calculated from the TV-image plotted in Figure 7(c). M_y curve has a sharp minimum approximately at $y = 90$ (row = 145, note that y is different than row number and equal to $y =$ 255-row) which corresponds to the location of the connected edges of vocal folds. This causes the separation of the glottis into two disjoint regions if the TV-image is thresholded by the sum of the mean and the standard deviation of the union of maxima curves. Since in the region selection step one of the regions is selected, calculated mask does not cover the glottis completely as shown in Figure 7(d). This problem can be solved by applying either one or both of the following methods:

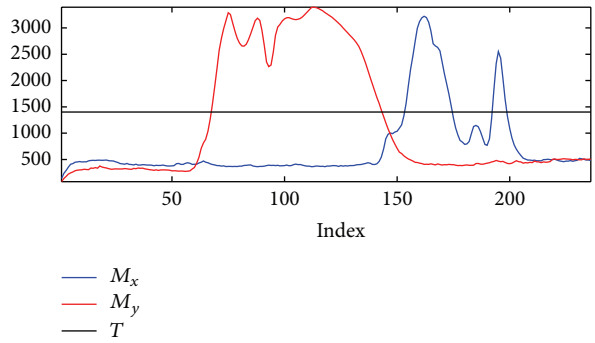
(1) thresholding TV-image by a smaller threshold, for example, $\mu+0.5\sigma$, or an adaptive method for selection of threshold from a set of candidates, that is, choosing a threshold from the set $T_i = \{\mu+0.1\sigma, \mu+0.2\sigma, \ldots, \mu+0.9\sigma, \mu+\sigma.\}$;

(2) 2D median filtering of TV-image.

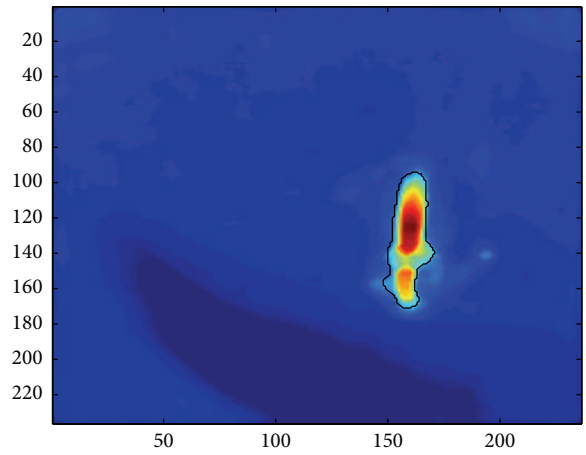
2D median filtering of TV-images is chosen as a general solution and used in the masking method. The TV-image obtained by 9×9 2D median filtering is shown in Figure 8(a). The TV of the glottal region is distributed over the location of the glottis after median filtering. Furthermore, some of the small regions having large TV resulting from shining of



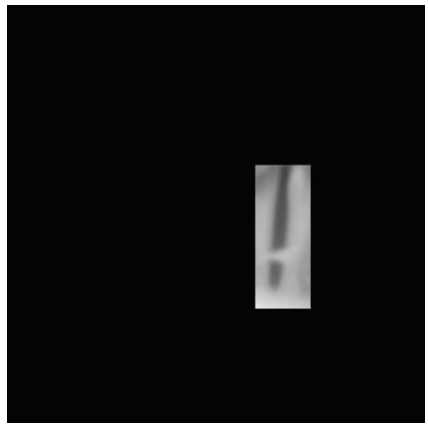
(a) TV-image after 2D median filtering



(b) M_x and M_y curves calculated from the TV-image in (a)

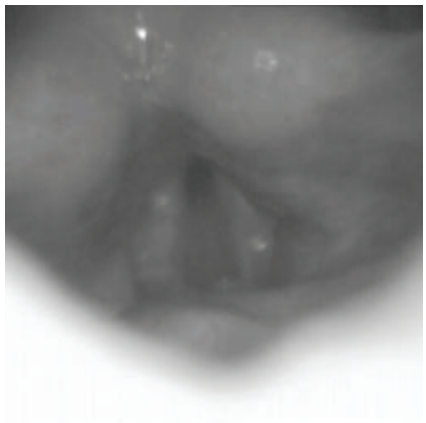


(c) The contour obtained after median filtering

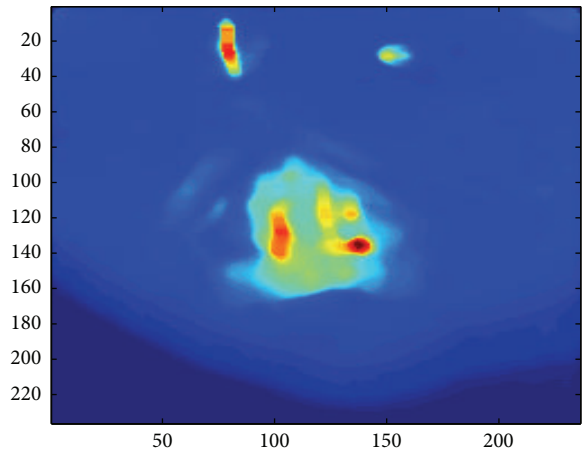


(d) Masked HSV after 2D median filtering of TV

Figure 8: Application of the proposed method with 2D median filtering when glottis is separated into two regions.



(a) A low contrast HSV snapshot



(b) TV-image of low contrast video in (a)

Figure 9: A low contrast HSV and its TV-image.

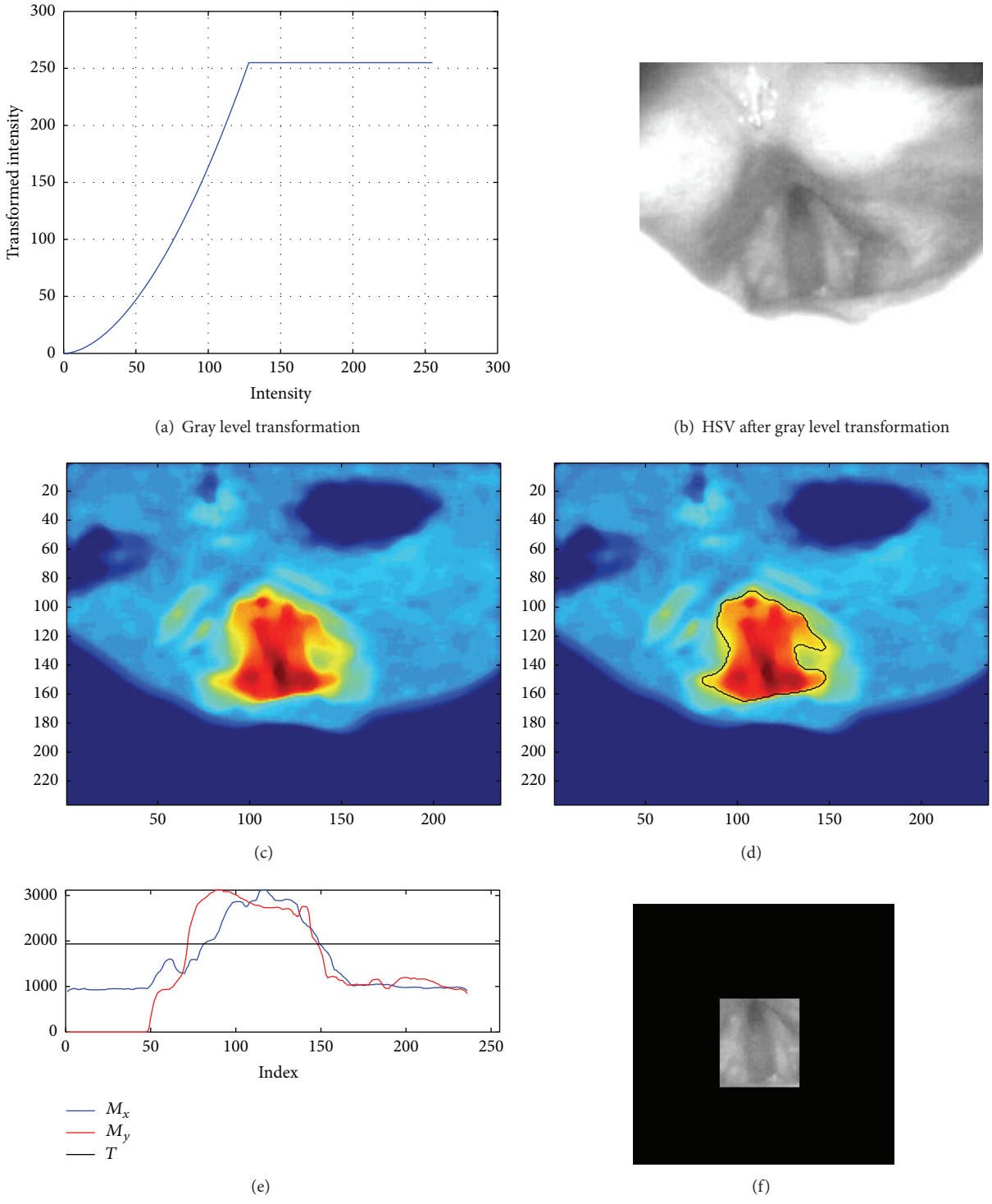


Figure 10: Application to a low contrast HSV.

tissues are diminished by the filtering (e.g., region around $x = 200$). Therefore, TV-image is enhanced by 2D median filtering. The maxima curves obtained from median filtered TV-image are plotted in Figure 8(b). Median filtering of the TV-image smooths the minimum in M_y (around $y = 90$) corresponding to the location of the connection of the two regions at the glottis. The contour of TV-image obtained after thresholding is shown in Figure 8(c). Now, both glottal regions are covered by the contour. The final constructed mask is shown in Figure 8(d). Without using 2D median filter, the masked HSV is erroneous as seen in Figure 7(d). However, by the use of median filtering on TV, the mask

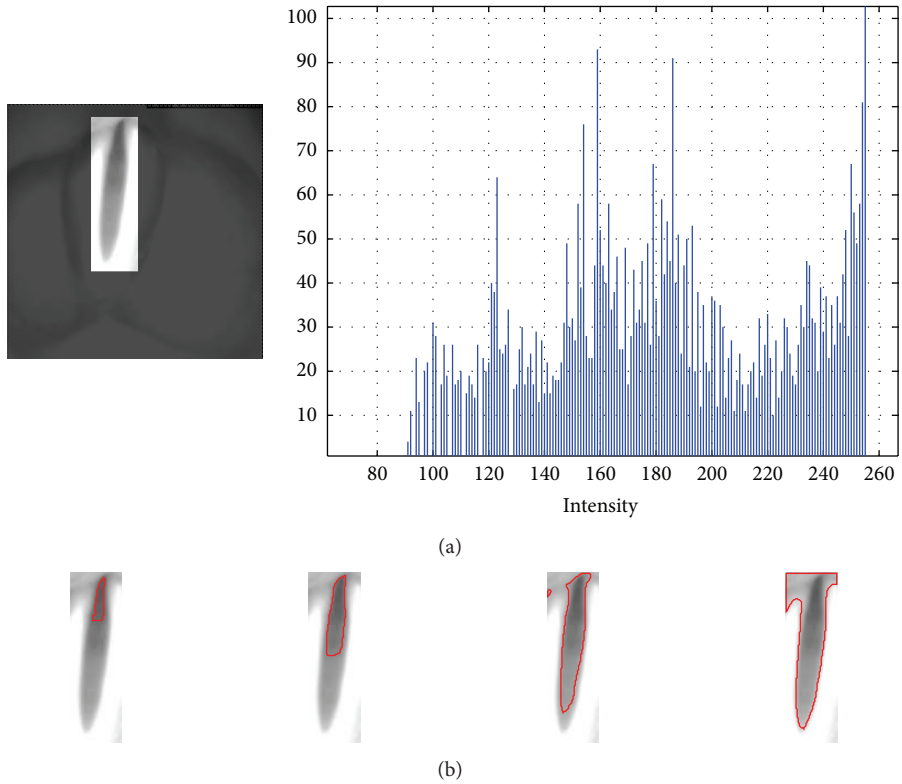


FIGURE 11: (a) Masked HSV image and its histogram. (b) Boundary of the vocal folds in the masked image obtained by thresholding with 110, 140, 170, and 210, respectively.

is corrected and the glottis is completely covered by the mask as plotted in Figure 8(d). In general, using 2D median filter enhances TV-images; hence, it is used in the automatic masking method.

Case 2 (enhancing low contrast images). Low contrast brings some difficulties in automatic masking HSV images. It requires a special approach which is described here. One low contrast HSV image is shown in Figure 9(a). The contrast of the image is relatively less than the previous HSV image examples. When the contrast of consecutive images is reduced, the total variation of intensity of pixels corresponding to moving objects decreases. Therefore, low contrast reduces TV values at the glottal activity region. The TV-image obtained from the image is shown in Figure 9(b). Note that there are two regions having significantly large TV value above the vocal folds. These regions result from the shining points above the vocal folds. The TV values of the glottis are high but distributed in an area about the location of the vocal folds. One solution is to use contrast transformation. The contrast of the HSV is transformed by the following nonlinear transformation:

$$T(I) = \begin{cases} 255 * \left(\dfrac{I}{I_{\text{High}}}\right)^{1.8}, & I \le I_{\text{High}}, \\ 255, & I > I_{\text{High}}, \end{cases} \tag{7}$$

where I is the intensity and I_{High} is the top 10% of all pixel values [13].

The transformation and transformed HSV are plotted in Figures 10(a) and 10(b). The transformation not only increases the contrast but also reduces the temporal intensity variations at the bright regions. The TV-image obtained after contrast transformation is shown in Figure 10(c). The transformation enhances the TV-image by reducing the effect of shining tissues due to illumination and increases the TV values of the active glottal region. The estimated contour and maxima curves obtained from the TV-image and resulting mask are shown in Figures 10(d) and 10(f), respectively. The method successfully locates the vocal folds.

3. Modeling Nonuniform Illumination

Histogram based image segmentation relies on the lighting conditions on the scene [13]. Under uniform lighting, or illumination, glottis and background in an HSV image can be separated using a single threshold determined by intensity histogram. However, the intensity distribution of the glottis and background is deformed under nonuniform illumination. It results in such a complex intensity distribution that separation of the glottis and background may not be achieved via simple thresholding. For instance, an HSV image distorted by nonuniform illumination and its histogram are shown in Figure 11(a). Note that, despite the successful

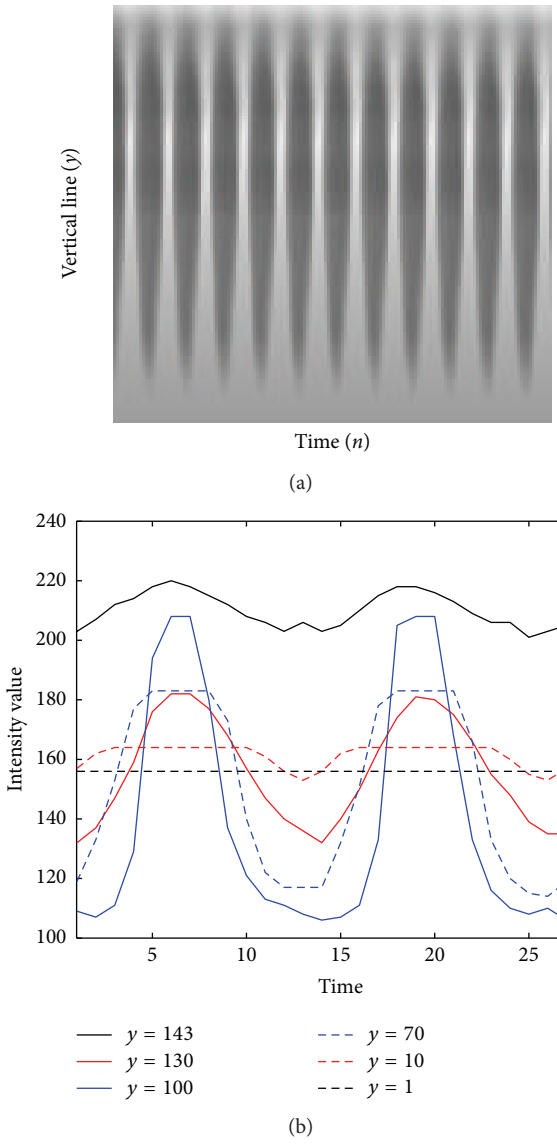


FIGURE 12: (a) Vertical slice image (VSI), intensity variation along the vertical slice of masked image. (b) Intensity variations of VSI image for different vertical lines.

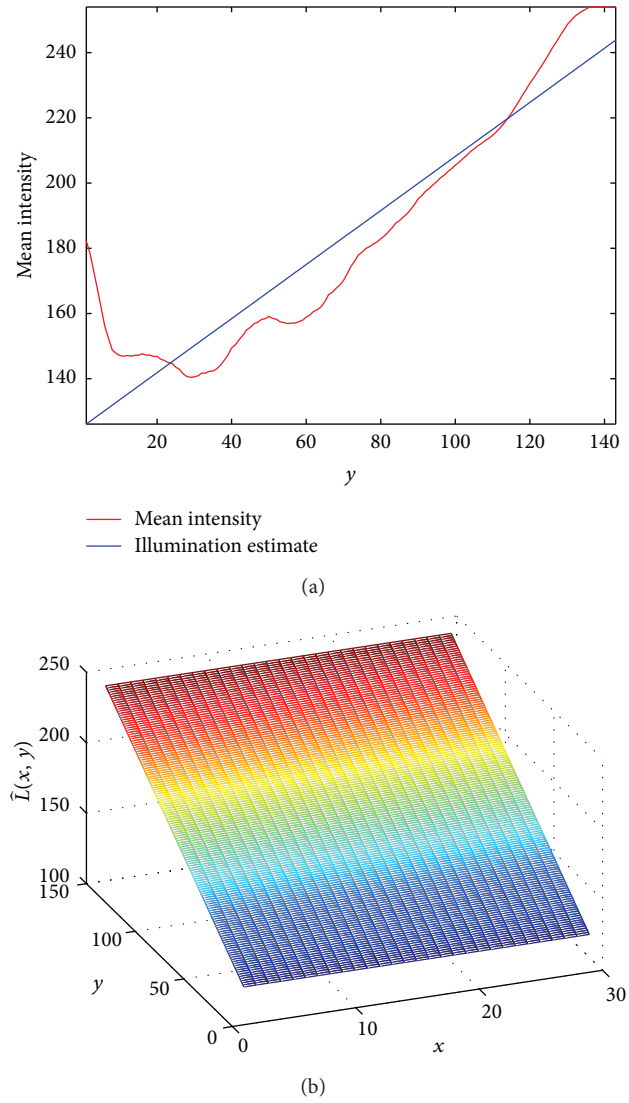
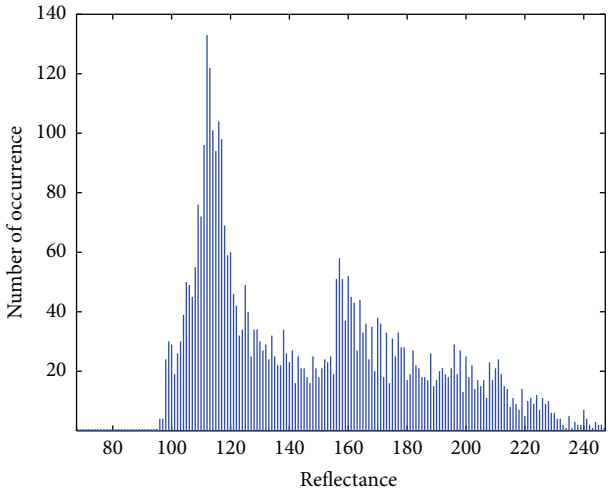


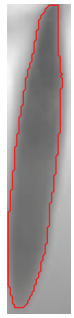
FIGURE 13: (a) Mean intensity along the center vertical line in masked HSV and estimated illumination line. (b) Estimated illumination function, $\widehat{L}(x, y)$.

masking over the image, the intensity distribution is far from having a bimodal character as indicated earlier in Section 2. A reliable and accurate threshold cannot be determined from the intensity histogram. For instance, the boundary of the vocal folds obtained by 4 different thresholds corresponding to the local minima of the histogram is shown in Figure 11(b). It is seen that the estimated boundaries are erroneous due to the deformation of the glottis and background intensity distributions.

In this section, we present a novel illumination model for HSV images and a method to estimate the parameters of the model. The method is based on the mean intensity distribution along the longitudinal cross section at the center of the glottis. Vertical slices of masked video frames are used to form, $\text{VSI}(n, y) = I_{\text{MASKED}}(x_{\text{center}}, y, n)$, where x_{center} is the horizontal center of the glottis when it is aligned vertically. The vertical slice image, $\text{VSI}(n, y)$, obtained from HSV frames plotted in Figure 11(a), is shown in Figure 12(a). It shows temporal intensity variation on the central vertical line of the glottis. The intensity on the vertical slice is low during the glottal opening and high when the folds are closed. Intensity variations for different y values of VSI are shown in Figure 12(b). For $y = 143$ (top line), the intensity is high and has a small variation in time with a small amplitude, and it has high periodicity. The largest variation is observed on $y = 100$ ($\text{VSI}(n, 100)$). It is the best representation of the periodic movement of the vocal fold edges. The variation is almost constant and exhibits nonperiodical movement on $y = 1$. This is due to the fact that the variation on $y = 1$ line does not contain information from the glottis region. The variation on the lines, $y = 130$ and $y = 70$, is also periodic with considerably large amplitudes. For $y = 10$, the amplitude variation is small. Note that the mean intensity



(a)



(b)

FIGURE 14: (a) Histogram of the reflectance image. (b) Estimated reflectance image and the boundary of the vocal folds obtained by thresholding with $T = 150$ (reflectance is scaled by 255 for grayscale representation).

values on the lines presented in the figure are different and depend on y. The temporal mean of the intensity variation on each horizontal line in the VSI image can be a measure of the illumination on that particular line. The temporal mean of the VSI as a function of y is plotted in Figure 13(a). Note that the mean intensity increases almost linearly from bottom ($y = 20$) to the top ($y = 143$) in the VSI.

The intensity of an image, $I(x, y)$, can be considered as a product of the illumination, $L(x, y)$, and reflectance, $R(x, y)$, the amount of illumination reflected from the objects in the scene ($I(x, y) = R(x, y) \cdot L(x, y)$) [13]. The illumination depends on the energy provided by the light source. The reflectance, $R(x, y)$, is the characteristic of the imaged objects and it is between 0 (total absorption) and 1 (total reflection). To extract the reflectance component, an estimate of the illumination function, $\widehat{L}(x, y)$, can be used.

Under uniform illumination ($L(x, y) = L$), intensity is expressed by reflectance times a constant ($I(x, y) = L \cdot R(x, y)$). Furthermore, if a uniformly reflective object ($R(x, y)$ is a constant, R) is imaged under uniform illumination, the intensity, $I(x, y)$, will be a constant. Similarly, if an image sequence is formed by periodic placement and removal of the object at the same location, the mean intensity variation along the image sequence at each point of the region covered by the object will be the same. As illustrated in this section, the intensities on vertical center line in the glottis are periodic. It is assumed that the vertical center line in the glottis is uniformly reflective ($R(y) = R$). It implies that, under uniform illumination, the mean intensity is the same at each point on the vertical center line. However, under nonuniform illumination, for the vertical center line, the mean intensity is the product of a constant reflectance, R, and the illumination function, $L(y)$. If R is chosen as 1 (for the sake of simplicity), the nonuniform illumination, $L(y)$, will be the same as the mean intensity on the vertical center line. In the same way, the illumination function, $L(x, y)$, can be estimated by temporally averaging the intensity for each pixel location in the glottis. To reduce the computational complexity, we assume that $L(x, y)$ changes only vertically and is the same on each horizontal line ($L(x, y) = L(y)$). The mean intensity is modeled by using a line as shown in Figure 13(a). The nonuniform illumination estimate, $\widehat{L}(x, y)$, is obtained by extending the line horizontally as shown in Figure 13(b). The reflectance component of the masked HSV image, $R(x, y)$, can be obtained by dividing the intensity of the masked image, $I(x, y)$, into the illumination estimate, $\widehat{L}(x, y)$. To represent the reflectance as a grayscaled image, its values are scaled to be confined to the interval [0–255] by multiplying 255. The estimated reflectance and its histogram are plotted in Figure 14.

The distinction between the glottis and background reflectance distribution can be observed from the clear bimodal structure in Figure 14(a). For discriminating the glottis from the background a reflectance threshold can be selected by visually inspecting the histogram. The reflectance value corresponding to the middle point between the peaks is found to be 150. It is used to convert the reflectance image into binary image and edge detection operation is applied on the binary image to extract the boundary of the vocal folds. Estimated boundary is illustrated in Figure 14(b). The vocal folds' edges are accurately determined by reflectance thresholding. Another nonuniform illumination modeling and reflectance based thresholding example is shown in Figure 15. The intensities of the image are distributed between values 120 and 255. Note that, due to high illumination, a large amount of intensity values is saturated around the upper intensity limit, 255. In this case, the classification of the glottis pixels based on the intensity distribution is problematic. A reliable threshold cannot be found even by visual inspection. However, the reflectance histogram obtained by using the proposed illumination model exhibits bimodal distribution. The result of the segmentation by using the reflectance value corresponding to the minimum between the two highest peaks in the reflectance histogram, $R = 130$, is shown in the upper right panel of the figure. The vocal folds' edges are accurately determined by using reflectance based thresholding.

4. GMM Based VSI Reflectance Modeling

One of the problems in the histogram based HSV processing is segmentation of small glottis regions appearing at the

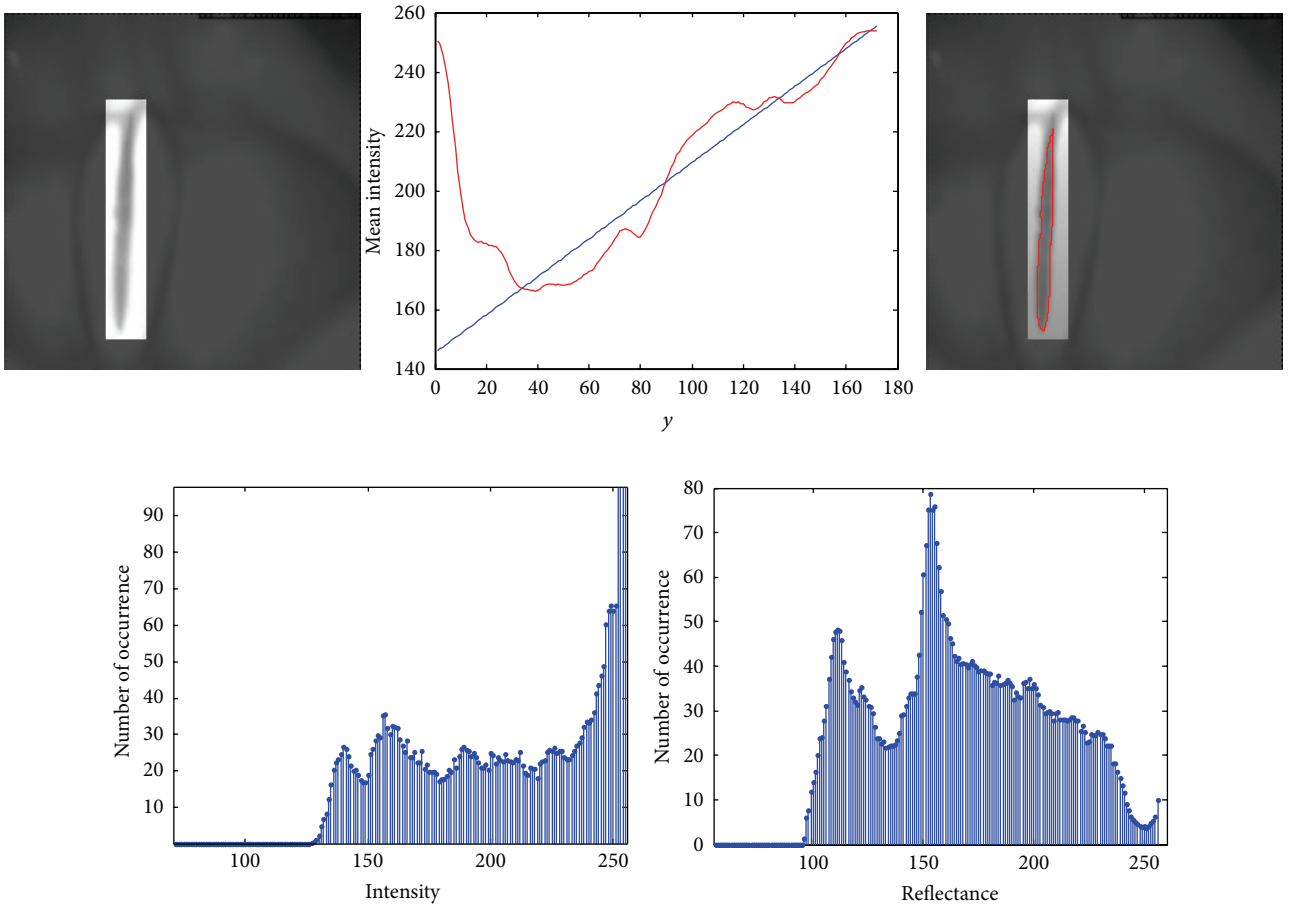


Figure 15: Illumination modeling system (reflectance is scaled by 255 for grayscale representation).

beginning of the glottal opening and at the end of the closing phases of vocal folds vibration. As illustrated in the introduction section, the intensity histograms of the images captured during the glottal closing or opening can be unimodal. It is not possible to estimate a reliable threshold even by manual intervention in those cases.

In this section, a novel automatic threshold determination method from consecutive HSV frames is presented. It is shown in the previous section that the reflectance histograms have better modality than the intensity histograms, due to the reduction of the effect of the nonuniform illumination. Therefore, reflectance based image segmentation is preferred in HSV segmentation. The VSI image based on reflectance represents reflectance distribution of the pixels belonging to the glottis and vocal fold edges. A reflection based VSI image is shown in Figure 16(a). The reflectance distribution provided by the vertical slice image is more reliable than that of the original unmasked image in distinguishing the boundary of glottal opening. The reflectance histograms of the unmasked, masked, and VSI images are shown in Figure 16(b). The unmasked image histogram is almost unimodal. Similarly, there is no clear boundary in the masked image histogram and it is not possible to separate the background and glottis easily. However, the histogram of the VSI image has clearly separated two peaks that lead to a robust threshold determination to separate glottis and background.

The VSI reflectance can be represented by a mixture of two Gaussian densities (GMM) [17]. The reflectance distributions at the glottal opening, $P_{R|\mathrm{GO}}(r \mid \mathrm{GO})$, and at other locations, $P_{R|\mathrm{O}}(r \mid \mathrm{O})$, form the two components. $P_{R|\mathrm{GO}}(r \mid \mathrm{GO})$, $P_{R|\mathrm{O}}(r \mid \mathrm{O})$, and $P(R)$ can be written as follows:

$$P_{R|\mathrm{GO}}(r \mid \mathrm{GO}) = N(r \mid \mu_1, \sigma_1) = \frac{1}{\sqrt{2\pi{\sigma_1}^2}} e^{(-1/2)((r-\mu_1)/\sigma_1)^2},$$

$$P_{R|\mathrm{O}}(r \mid \mathrm{O}) = N(r \mid \mu_2, \sigma_2) = \frac{1}{\sqrt{2\pi{\sigma_2}^2}} e^{(-1/2)((r-\mu_2)/\sigma_2)^2},$$

$$P(R) = \sum_{i=1}^{2} w_i N(r \mid \mu_i, \sigma_i), \tag{8}$$

where w_i's are weights of the Gaussian components. The parameters of the GMM, μ_i, σ_i, and w_i, are estimated by EM (expectation and maximization) algorithm [18]. The posterior probabilities, $P_{\mathrm{GO}|R}(\mathrm{GO} \mid r)$ and $P_{\mathrm{O}|R}(\mathrm{O} \mid r)$, are written as follows:

$$P(i \mid R) = \frac{w_i N(R \mid \mu_i, \sigma_i)}{\sum_{k=1}^{2} w_k N(R \mid \mu_k, \sigma_k)}, \tag{9}$$

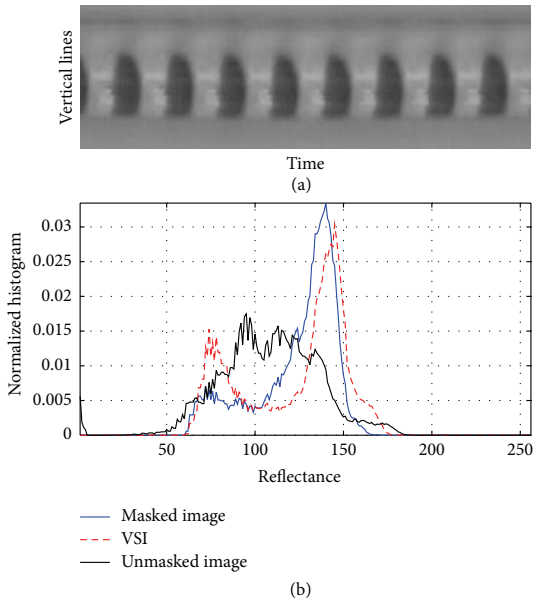
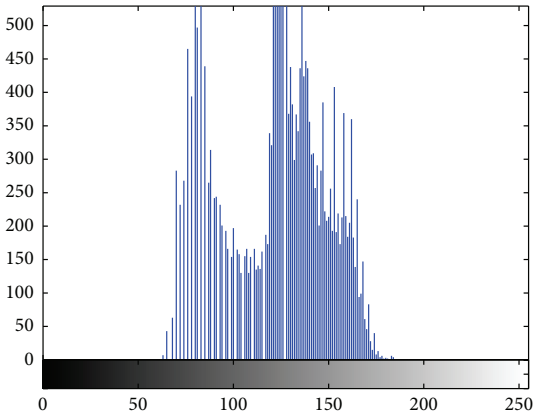
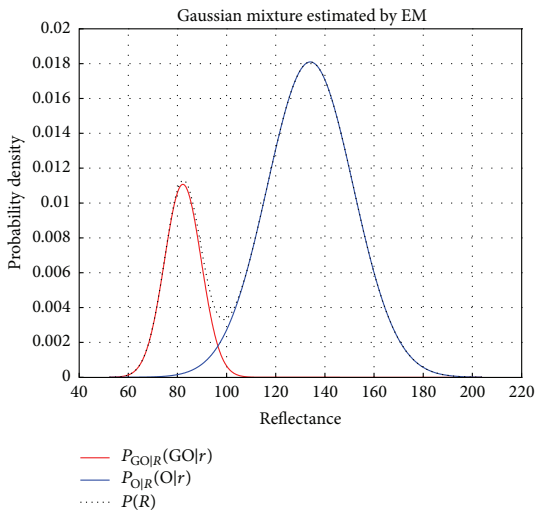


FIGURE 16: (a) Vertical slice image (VSI). (b) Reflectance histograms of unmasked, masked, and VSI images (reflectance is scaled by 255 for grayscale representation).



(a) Reflectance histogram



(b) Estimated GMM by EM algorithm

FIGURE 17: Estimation of GMM.

where $i = 1, 2$ denotes glottal opening (GO) and others (O), respectively. A VSI histogram and corresponding GMM are shown in Figure 17. Estimated GMM fits the reflectance distribution well.

The estimated densities are used to segment each HSV frame in the image sequence as

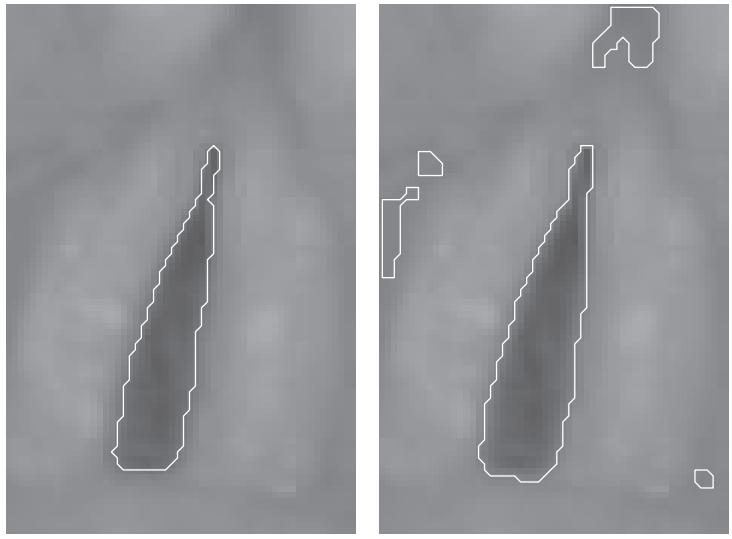
$$g(x, y, n) = \begin{cases} 1, & P(\mathrm{GO} \mid R(x, y, n)) > P(\mathrm{O} \mid R(x, y, n)), \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

In the last step of the algorithm, the glottal area variation in HSV is determined. The glottal area in terms of the number of pixels is calculated as

$$a_g[n] = \sum_{x,y \in \mathrm{Mask}} g(x, y, n). \tag{11}$$

The steps of the algorithm are given as follows.

(1) Get N images from HSV.
(2) *Start automatic mask determination.*
(3) Calculate TV-image (see (1)).
(4) Apply KxK 2D median filter to TV-image (the smaller K, the sharper TV-image).
(5) Calculate M_x and M_y (see (2)).
(6) Threshold for segmentation of TV-image is determined by the union of M_x and M_y statistics, $+\sigma$.
(7) Segment TV-image and find the regions inside it using connected component method.
(8) Select the region having largest E (see (3) and (4)) as glottis and determine the boundary of the region by edge detection.
(9) Construct a mask by using the extrema of the boundary (see (6)).
(10) *End masking.*
(11) *Start illumination modeling.*
(12) Apply masking to image sequence and form VSI image ($\mathrm{VSI}(n, y) = I_{\mathrm{MASKED}}(x_{\mathrm{center}}, y, n)$).
(13) Calculate the mean intensity at each row in VSI and estimate illumination function $\widehat{L}(x, y)$.
(14) Estimate reflectance images from masked images by using $R(x, y) = I(x, y)/\widehat{L}(x, y)$.
(15) *End illumination modeling.*
(16) *Start density estimation.*
(17) Form VSI image of the reflectance image sequence.
(18) Estimate GMM parameters by EM method and calculate the posterior probabilities, $P_{\mathrm{GO}|R}(\mathrm{GO} \mid r)$ and $P_{\mathrm{O}|R}(\mathrm{O} \mid r)$ (see (8)–(9)).
(19) *End density estimation.*
(20) *Start segmentation.*
(21) Segment each image by using the posterior probabilities (see (10)).
(22) Calculate glottal area (see (11)).
(23) *End segmentation.*



(a) (b)

Figure 18: Segmentation results. (a) Proposed method. (b) The method of [7].

5. Results

Proposed algorithm is compared with the method of [7] on IRCAM HSV database [19]. In [7], intensities in HSV frames are modeled by Rayleigh distribution; then a threshold is determined by a Bayesian approach. After thresholding, region-growing operation is applied. In this study, region growing is omitted, because it is optional and its success depends on the accurate determination of the glottis region. The core function of [7] is Rayleigh based intensity modeling and thresholding is used in the comparison. However, it is our experience that the application of the method of [7] on unmasked frames frequently produces undesirable segmentation results. It is expected since unmasked images usually have unimodal intensity distribution. Therefore, the method of [7] is applied on the masked images obtained from the first stage of our algorithm. Examples of segmentation results by the proposed method and the method of [7] are shown in Figure 18. Segmented regions other than glottis are observed in the result obtained by the method of [7]. On the other hand, the use of masking and VSI image improves the success of classification significantly. For comparison, 3000 images are segmented from 10 different HSV videos (300 for each) manually to produce a reference data set. Segmentation results for three different cases are shown in Figures 19–21. The glottal area waveforms obtained by manual segmentation, by the method of [7] and the proposed method are shown in Figure 22. Mean square error (MSE) values of the glottal area waveforms obtained by applying the two methods are given in Table 1.

According to the results presented in Table 1, the MSE of the proposed algorithm is smaller than the MSE of the method of [7] by about 94%. The accuracy in the glottal area and vocal fold boundary estimates is increased significantly.

Table 1: MSEs over 3000 frames of the estimated glottal area waveforms.

	Proposed	The method of [7]	Decrease in MSE (%)
MSE	5969.9	93573	93.60

6. Conclusion

Automatic segmentation of HSV images is necessary for investigation of vocal folds' functions in clinical and speech production studies. In this study, an automatic method is introduced for extracting glottal area waveform from HSV images. The first novelty in the method is that the region of interest is determined automatically by a robust masking algorithm using TV-norm of consecutive images which produces a strong indication for moving vocal folds. Using a sliding HSV block, it can be updated easily for processing long sequence of frames. A planar illumination model for HSV images is the second unique feature of the method. By means of the proposed planar illumination model, reflectance images can be obtained and used in the vocal fold segmentation. The other novelty is, by introducing vertical slice image, the use of reflectance variation information in the glottis from multiple images instead of using a single image to model the reflectance distribution. Modeling the reflectance distribution of vertical slice images by GMM is easy, computationally efficient and produces more accurate glottal area waveforms. For long image sequences, the mean and variance of GMM can be recursively updated by forming a sliding frame sequence framework in constructing the VSI image. The proposed approach can be used in any histogram based automatic image segmentation systems.

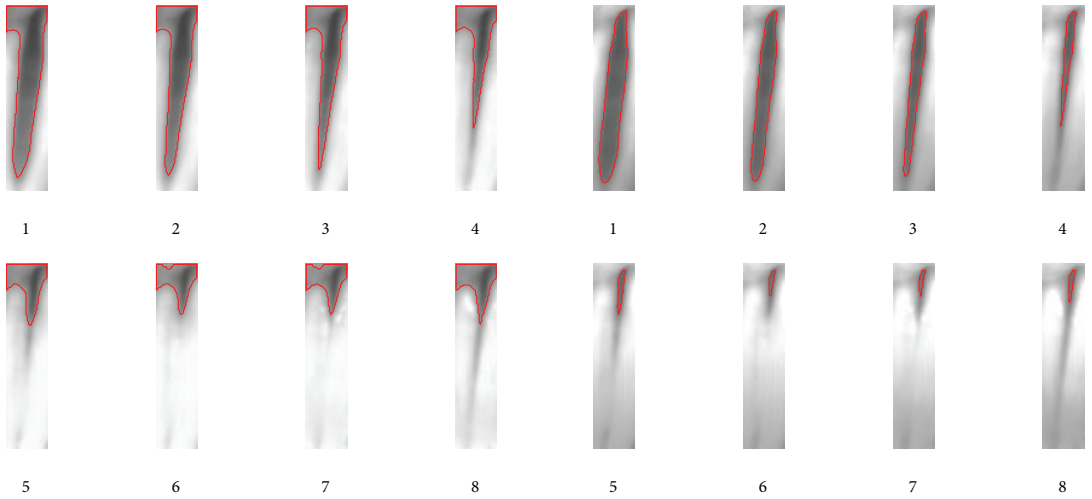


Figure 19: Case 1: Segmented HSV image sequence by the method of [7] (a), the proposed method (b).

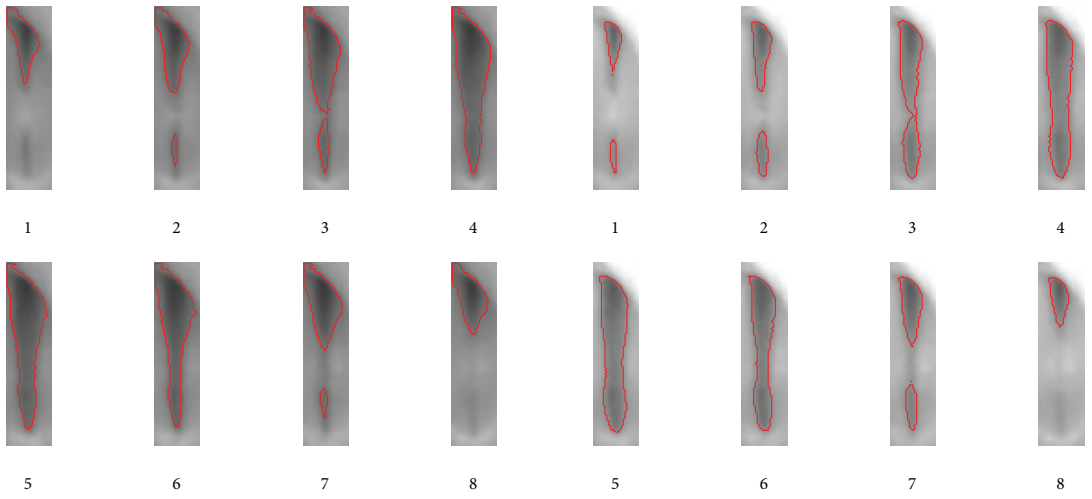


Figure 20: Case 2: Segmented HSV image sequence by the method of [7] (a), the proposed method (b).

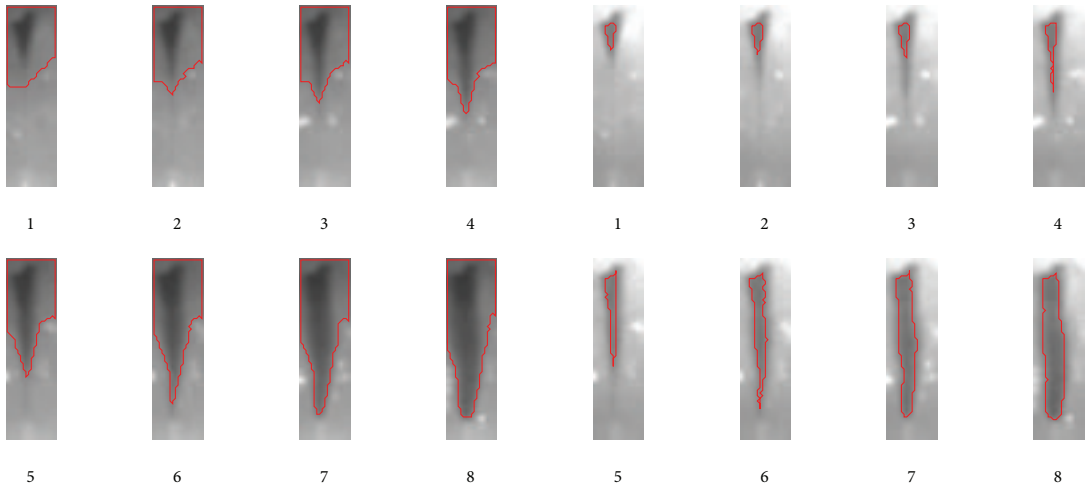


Figure 21: Case 3: Segmented HSV image sequence by the method of [7] (a), the proposed method (b).

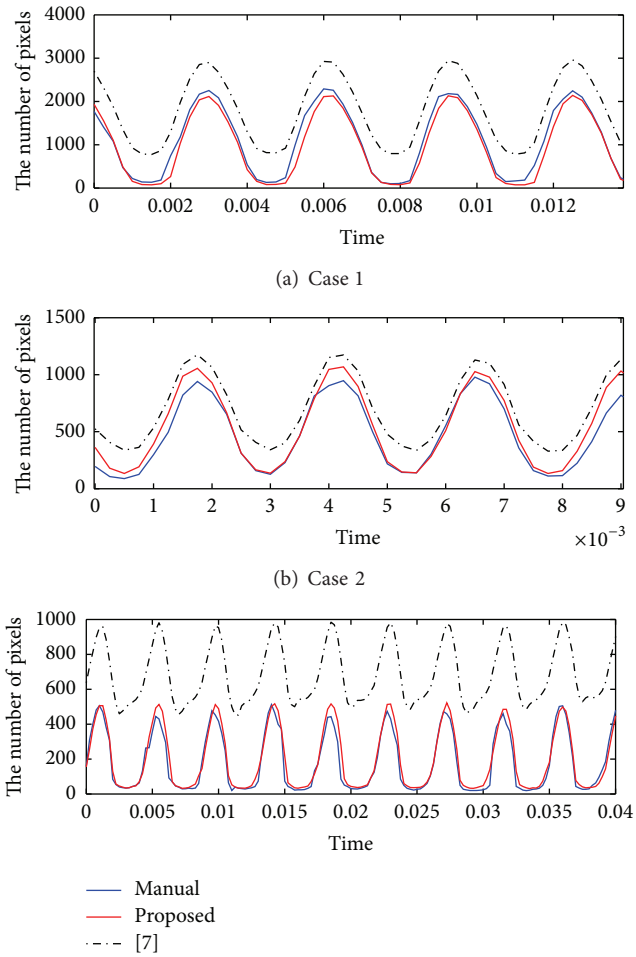


FIGURE 22: Manually extracted and automatically estimated glottal area signals Cases 1–3.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

Special thanks are due to Gilles Degottex and Erkki Bianco for sharing their valuable IRCAM HSV database.

References

[1] D. D. Mehta, D. D. Deliyski, S. M. Zeitels, T. F. Quatieri, and R. E. Hillman, "Voice production mechanisms following phonosurgical treatment of early glottic cancer," *Annals of Otology, Rhinology and Laryngology*, vol. 119, no. 1, pp. 1–9, 2010.

[2] Y. Yan, K. Ahmad, M. Kunduk, and D. Bless, "Analysis of vocal-fold vibrations from high-speed laryngeal images using a Hilbert transform-based methodology," *Journal of Voice*, vol. 19, no. 2, pp. 161–175, 2005.

[3] A. Verikas, A. Gelzinis, D. Valincius, M. Bacauskiene, and V. Uloza, "Multiple feature sets based categorization of laryngeal images," *Computer Methods and Programs in Biomedicine*, vol. 85, no. 3, pp. 257–266, 2007.

[4] J. Unger, M. Schuster, D. J. Hecker, B. Schick, and J. Lohscheller, "A multiscale product approach for an automatic classification of voice disorders from endoscopic high-speed videos," in *Proceedings of the 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC '13)*, IEEE, 2013.

[5] C. Tao, Y. Zhang, and J. J. Jiang, "Extracting physiologically relevant parameters of vocal folds from high-speed video image series," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 5, pp. 794–801, 2007.

[6] X. Qin, S. Wang, and M. Wan, "Improving reliability and accuracy of vibration parameters of vocal folds based on high-speed video and electroglottography," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 6, pp. 1744–1754, 2009.

[7] Y. Yan, X. Chen, and D. Bless, "Automatic tracing of vocal-fold motion from high-speed digital images," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 7, pp. 1394–1400, 2006.

[8] Y. Yan, G. Du, C. Zhu, and G. Marriott, "Snake based automatic tracing of vocal-fold motion from high-speed digital images," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '12)*, IEEE, 2012.

[9] T. Ikuma, M. Kunduk, and A. J. McWhorter, "Preprocessing techniques for high-speed videoendoscopy analysis," *Journal of Voice*, vol. 27, no. 4, pp. 500–505, 2013.

[10] D. D. Mehta, D. D. Deliyski, T. F. Quatieri, and R. E. Hillman, "Automated measurement of vocal fold vibratory asymmetry from high-speed videoendoscopy recordings," *Journal of Speech, Language, and Hearing Research*, vol. 54, no. 1, pp. 47–54, 2011.

[11] 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*, vol. 11, no. 4, pp. 400–413, 2007.

[12] 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*, vol. 54, no. 5, pp. 641–654, 2012.

[13] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, Prentice Hall, 2002.

[14] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena*, vol. 60, no. 1–4, pp. 259–268, 1992.

[15] C. R. Vogel and M. E. Oman, "Iterative methods for total variation denoising," *SIAM Journal on Scientific Computing*, vol. 17, no. 1, pp. 227–238, 1996.

[16] T. F. Chan and C.-K. Wong, "Total variation blind deconvolution," *IEEE Transactions on Image Processing*, vol. 7, no. 3, pp. 370–375, 1998.

[17] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley-Interscience, New York, NY, USA, 2nd edition, 2001.

[18] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society B. Methodological*, vol. 39, no. 1, pp. 1–38, 1977.

[19] G. Degottex and E. Bianco, "IRCAM Databases of High Speed Videoendoscopy," UPMC-Ircam, France, 2010.