Graphical Evaluation of Vocal Fold Vibratory Patterns by High-Speed Videolaryngoscopy

***Alan P. Pinheiro, †Maria Eugênia Dajer, †Adriana Hachiya, ‡Arlindo N. Montagnoli, and †Domingos Tsuji,** **Patos de Minas, Minas Gerais, †São Paulo, and ‡São Carlos, Brazil*

Summary: Objective. To characterize the voice and vocal fold function of an individual, it is essential to evaluate vocal fold vibration. The most widely used method for this purpose has been videolaryngoscopy.
Methods. This article proposes a digital image processing algorithm to estimate the glottal area (ie, the space between the vocal folds) and produce graphs of the opening and closing phases of the glottal cycle. In eight subjects without voice disorders, vocal fold movements were recorded by high-speed videolaryngoscopy at 4000 frames per second. The video data were processed by a combination of image segmentation techniques that estimate the glottal area. The segmented area was used to construct the glottal waveform.
Results. The graphs revealed important properties of vocal fold vibration, including amplitude, velocity, and other characteristics that have a major influence on voice quality.
Conclusions. The combination of the high-speed technology with the proposed method improves the vocal fold analysis given a numerical feedback through graphical representation of the real vibratory patterns of the folds.
Key Words: Glottal area–Segmentation–Videolaryngoscopy–High-speed imaging.

INTRODUCTION

During phonation, the vocal folds produce self-sustained oscillation that models the airflow from the lungs, generating a primary signal that is used as a source of excitation of the vocal tract. This signal is known as the glottal flow and determines some of the basic properties of the voice, including its fundamental frequency and main spectral components, which are modeled in the vocal tract[1] and allow the production of different sounds.

Typically, many voice pathologies are caused by changes in the structure or physiology of the vocal folds and such changes reflect in the glottal flow waveform.[2] Therefore, it is important to develop methods for studying the vocal folds. Various methods have been proposed to analyze the vocal folds and the glottal flow waveform. Such methods include inverse filtering of the signals,[3] electroglottography,[4] and videolaryngoscopy.[5]

In clinical and research settings, the gold standard for the analysis of the vocal folds and their vibratory motion is videolaryngoscopy.[6] However, videolaryngoscopy is generally performed with cameras that can record no more than 20–30 frames per second (fps). Because the vocal folds vibrate at a mean frequency of 100–400 cycles per second, it is impossible to record their real vibratory motion by means of conventional laryngoscopy. Therefore, traditional videolaryngoscopy is used in an attempt to record some samples of phases of the vibratory cycle of the folds and create an artificial cycle using stroboscopic techniques. Because the videolaryngoscopy provides images that lack detail, it is most frequently used exclusively for qualitative evaluations.

The advent of high-speed cameras and the possibility of using such cameras in conjunction with videolaryngoscopy opened new avenues for clinical and scientific research on vocal function. High-speed videolaryngoscopy provides detailed images of the entire cycle of vocal fold vibration. As an immediate result of this new technology, various studies have been conducted to quantify physiological vocal fold parameters,[7–9] investigate the aerodynamic transfer of energy from glottal airflow to vocal fold tissue during phonation,[10] determine the mechanical properties of vocal fold tissues,[11] and thoroughly evaluate the effects of specific voice disorders.[2] Studies comparing conventional with high-speed videolaryngoscopy have confirmed the advantages of the latter over the former.[5,6]

High-speed imaging techniques can provide a large quantity of images (approximately 2000–4000 fps). Because manual analysis of such images would be tedious, new digital image processing methods are required. In addition, high-speed videolaryngoscopy produces low-resolution images affected by noise that is inherent to the imaging method as well as producing uneven illumination. Studies aimed at developing new algorithms for processing such images have, therefore, been conducted.[12–14] However, most of the methods presented in those studies showed little or no automaticity, having been used to analyze excised vocal fold images obtained under ideal laboratory conditions and being dependent on good image resolution.

The objective of the present study was to develop a computational method for calculating (segmenting) the glottal area (ie, the space between the vocal folds) using images acquired by high-speed videolaryngoscopy. By calculating the glottal area, it is possible to estimate the glottal flow waveform that excites the vocal tract and characterize the opening and closing of the vocal folds. The present study focused on the application of the method in the evaluation of *in vivo* clinical images, which often contain artifacts. In addition, the performance and sensitivity of the method were tested to determine its validity and robustness.

Accepted for publication July 30, 2013.
The present study received financial support from the São Paulo Research Foundation (FAPESP; grant no. 2010/18488).
From the *Nucleus of Scientific and Technological Development, Faculty of Electrical Engineering, Federal University of Uberlândia, Patos de Minas, Brazil; †School of Medicine, University of São Paulo, São Paulo, Brazil; and the ‡Department of Electrical Engineering, Federal University of São Carlos, São Carlos, Brazil
Address correspondence and reprint requests to Alan P. Pinheiro, Universidade Federal de Uberlândia, Av Getúlio Vargas 230, Patos de Minas, Minas Gerais, CEP 38.700-128, Brazil. E-mail: alan@eletrica.ufu.br
Journal of Voice, Vol. 28, No. 1, pp. 106-111
0892-1997/$36.00

http://dx.doi.org/10.1016/j.jvoice.2013.07.014

METHODS

Data collection

Eight individuals (four males and four females), with healthy voices, were invited to participate in the present study. The mean age was 42 years. All the participants underwent videolaryngoscopy with a high-speed camera (ENDOCAM 5262; Richard Wolf, Knittlingen, Germany), which was capable of recording 4000 fps for a duration of 2 seconds. The images had a resolution of 256×256 pixels. Figure 1 shows one of the study participants undergoing high-speed videolaryngoscopy. For high-speed videolaryngoscopy, all the participants were asked to sustain the vowel /a/. The study procedures were approved by the Research Ethics Committee of the University of São Paulo, School of Medicine (protocol no. 0767/09).

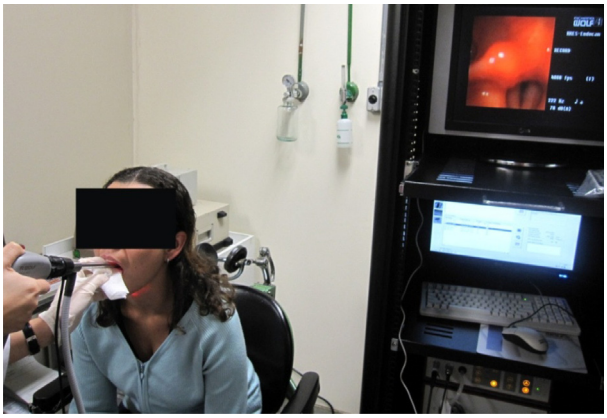

FIGURE 1. Videolaryngoscopy with a high-speed camera.

Glottal area segmentation

The algorithm developed to estimate (ie, segment) the glottal area as seen on those images consists of a sequence of steps that are applied to each frame. Figure 2 shows a block diagram illustrating the procedure. Unlike other methods,[12–14] which are aimed at segmenting the glottal area directly in few computational steps, the method developed in the present study was to obtain an initial rough estimate of the location of the glottal area and refine that estimate in subsequent steps until the entire space between the right and left vocal folds has been defined as accurately as possible.

The first step of the method was designated initialization (Figure 2A), in which the first frame of the high-speed video was displayed. In the initialization step, the user should select n points ($n \geq 1$) using the computer mouse. The selected points should be in the glottic region and are used to estimate the mean color intensity (μ) in the glottic region as shown in Equation (1), where i and j indicate the image coordinates of the selected points. This is the only manual step of the method, being performed only once for each examination.

$$\mu = 1/\mathrm{n} \sum_{x=1}^{\mathrm{n}} M([i,j]). \tag{1}$$

The second step of the method was designated binarization (or thresholding), in which the glottic region was separated from the remaining regions seen on the image on the basis of the threshold, which was calculated by the Equation (2).

$$\begin{aligned} T &= \mu \pm \sigma(i,j) \\ \sigma(i,j) &= \sqrt{\frac{1}{x \cdot y} \sum_{i=1}^{x} \sum_{j=1}^{y} (M[i,j] - \mu)^2}. \end{aligned} \tag{2}$$

In Equation (2), T indicates the threshold; μ the mean color intensity in the glottic region calculated in Equation (1); and σ the standard deviation of the surroundings of each pixel in the image that will be binarized.

The method used a technique known as local adaptive thresholding, which is based on the statistical relationship between each pixel and its surroundings. The algorithm uses the mean color intensity in the glottic region, as defined by the points selected manually in the first step, and the standard deviation of the surroundings of each pixel in the image that will be binarized. If the color intensity of the pixel at the coordinate point is lower than the threshold, the pixel is displayed in white, whereas the remaining (ie, unselected) pixels are displayed in black. Therefore, the regions of the image in which the color is similar to that of the glottal area defined in the initial step are selected, the remaining regions being eliminated from the image. Preliminary analyses reveal that most of the glottis was identified and displayed in white (Figure 2B). However, due to lighting problems and low contrast, certain regions outside the glottis were selected because their color was similar to that of the glottal area. Those selections constitute typical noises in the image.

The third step of the method was designated as seed estimation (Figure 2C), in which the image that was binarized in the previous step was filtered by means of a traditional method of image erosion and dilation,[15] thereby eliminating small noises (seen in white in Figure 2B) and irrelevant details. After the image had been filtered, a blob algorithm[16] was applied to the image. The algorithm identifies the largest segmented element (named blob) in the binary image and classifies it as being within the target area (the glottal area, in this case). Finally, the centroid of that blob was estimated, and the coordinates of the centroid were used as a "seed" in the next step. The seed point estimation process is used to automatically estimate, with some degree of certainty, a single coordinate on the image within the glottic region on the basis of the binary image. As previously mentioned, the user manually selects points in the glottic region only on the first frame of the video. For the remaining frames, the estimation is automatically provided by seed estimation. Cases in which the glottal area is small or absent (eg, cases in which the vocal folds are closed), the image erosion process eliminates all image elements and no seed is defined, the area, therefore, being classified as null.

After seed estimation, the region growing step is initiated (Figure 2D). The technique used in the present study expands the seed iteratively (generating a segmented area) by analyzing all pixels that surround the boundary of this segmented area in accordance with a homogeneity criterion that evaluates the difference between the intensity value of the surrounding pixels and the mean intensity of the segmented area.[17] If that