Quantitative Detection of Substitute Voice Generator During Phonation in Patients Undergoing Laryngectomy

Maria Schuster, MD; Frank Rosanowski, PhD, MD; Raphael Schwarz; Ulrich Eysholdt, PhD, MD; Jörg Lohscheller, PhD

Objective: To evaluate the vibration pattern of the substitute voice generator of patients who have undergone laryngectomy. For automatic quantification of the oscillations of the pharyngoesophageal (PE) segments, image processing of digital high-speed video sequences is applied.

Design: Physiologic analysis.

Setting: An acute care hospital.

Patients: Endoscopic recordings were taken of 10 men who underwent laryngectomy (mean ± SD age, 61.5 ± 5.2 years) during sustained phonation of a vowel using a 90° endoscope coupled to a high-speed camera.

Main Outcome Measures: An image-processing algorithm was developed to automatically define the pseudoglottis in each recording and track its movements.

Results: The clinical assessment of the high-speed technique for the endoscopic examination of the substitute voice generator yields the following results. The forms and oscillation characteristics of the pseudoglottides varied considerably: 3 pseudoglottides were circular, 6 were split shaped, and 1 was triangle shaped. A quasi-periodic opening and closing were observed and automatically detected by the described algorithm in each recording independently from quality of the recording and from morphologic and oscillation characteristics of the PE segment. The frequencies of the extracted oscillations of the pseudoglottides correspond to the structure of the acoustic signals.

Conclusions: Automatic image processing of PE segments derived from high-speed endoscopic recordings enables the detection and quantification of the substitute voice generator's oscillations in high temporal resolution. These data directly prove that the detected pseudoglottis is the source of the substitute voice. Close relations between substitute voice and functional properties of the PE segment exist. In the future, these data will be interpreted by applying biomechanical models of the PE segment. Presumably, results may help to optimize surgical and adaptive procedures for specific substitute voice restoration.

Arch Otolaryngol Head Neck Surg. 2005;131:945-952

Several functional limitations result from total laryngectomy, with the loss of voice being of major importance. Voice can be restored by an external sound generator (ie, an electrolarynx), which will not be discussed in this study, or by an internal sound generator in esophageal and tracheoesophageal speech. The sound generator of both esophageal and tracheoesophageal speech is the mucosa of the pharyngoesophageal (PE) segment. The PE segments differ individually, depending on the shape and stiffness of the scar between the hypopharynx and esophagus, the localization of the carcinoma, different surgical needs and procedures, and the extent of remaining esophageal mucosa. Although the tracheoesophageal voice resembles the normal voice¹ and is therefore commonly preferred in clinical practice, the results after voice restoration and the patients' estimation of substitute voice skills differ.² Several investigations of the substitute voice attempted to detect correlations between substitute voice quality and morphologic or dynamic properties of the PE segment. Mainly, these studies²⁻¹³ described the substitute voice by analyzing the acoustic signals, determination of aerodynamic parameters, radiologic investigations with contrast agents, and endoscopic techniques. However, because these techniques had only low temporal resolution and usually lacked a combined assessment of both morphologic and functional properties, dynamic properties of the sound-generating PE segment were not sufficiently detectable. Because of the vary-
The introduction of the high-speed imaging technique into endoscopy was a reasonable advance in the examination of the substitute voice generator: dynamic procedures can be recorded with a time resolution of approximately 4000 pictures per second. Originally, this method was developed for the examination of laryngeal voices and applied to several laryngeal physiologic and pathologic features. Hence, the vibrations of the PE segment during phonation can also be visualized in real time. Therefore, even slight and irregular movements can be seen. However, up to now the description of the oscillations has still been restricted. For this reason, we developed an automatic analyzing technique to obtain quantitative information on the vibration pattern of the PE segment. This technique should enable the detection of the PE segment independently from its location in an image, its shape, its vibration pattern, and other moving structures near the PE segment in a sequence, and it will track the contour of the PE segments during vibration in a high temporal resolution. Its contribution to substitute voice research will be exemplified by 10 patients who underwent laryngectomies.

**METHODS**

**PATIENTS**

Ten male speakers (L1-L10; “L” indicates laryngectomized individual) who had undergone laryngectomy (mean ± SD age, 61.5 ± 5.2 years) took part in this study after signing informed consent forms. The investigation was approved by the Ethics Review Committee of University Hospital. The patients had been using a voice prosthesis device for at least 12 months. Total laryngectomy was performed because of laryngeal cancer in 7 patients and hypopharyngeal cancer in 3 patients 1 to 7 years (mean ± SD, 4.2 ± 2.0 years) before the examination. Following surgical therapy, all of the patients underwent radiotherapy. None had any signs of relapse or metastases.

**HIGH-SPEED IMAGING**

The oscillating PE segments were endoscopically recorded during phonation of the vowel /a/ (as in bag) at a comfortable pitch while a high-speed camera system was coupled to a rigid 90° endoscope with an outer diameter of 9 mm (Wolf Corp, Knittlingen, Germany) (Figure 1). The high-speed camera systems were prototypes of the High-Speed Endocam by Wolf Corp and consisted of a camera with a frame rate of 3704 or 4000 Hz, respectively, and a charge-coupled device sensor of 64 × 128 pixel spatial resolution (discretization, 8 bit) or 256 × 128 pixel (discretization, 6 bit), respectively. A 250-W xenon-light lamp served as the light source. The acoustic signal was recorded simultaneously during phonation (B&K microphone, model 4129; Brüel & Kjaer Corp, Naerum, Denmark; sampling rate, 44.4 kHz; discretization, 16 or 8 bit, respectively).

**IMAGE PROCESSING**

The quantitative description of PE dynamics requires a complex image-processing procedure that identifies characteristic features of the pseudoglottis (area of the PE segment opening) within an image sequence. Automatic analysis of the vibrations of the pseudoglottis is performed by a 2-step procedure that tracks the deformations of the PE segment within a high-speed sequence by (a) automatic detection of the region of interest (ROI) (the region where the PE segment is located) and (b) contour tracking of the pseudoglottis. The first step (step a) is an initialization procedure that detects the location of the PE segment but does not detect its shape. It identifies specific and distinguishable features of the pseudoglottis in a gray-value picture, orienting on pixel intensity distribution, edges, and geometric properties to formulate mathematical criteria. The better that specific features can be identified, the easier a mathematical formulation of pseudoglottis criteria can be found. For the pseudoglottis, morphologic differences make it difficult to identify common geometric features. Therefore, additional information is necessary and can be acquired by fusing multiple sensor data, in this study the...
optic and acoustic sensor data, to detect the ROI within a high-speed sequence. Three criteria for PE segment detection can be formulated. First, the time-dependent pixel intensity correlates with the acoustic signal if the pixel belongs to the image region where the process of sound production takes place. Second, the opening and closing of the PE segment cause a change of the pixel intensity. Third, pixels with a low mean intensity represent the pseudoglottis. Fourth, the closer a pixel's localization to the center of the pseudoglottis, the higher its contribution to the substitute voice–generating process. The combination of the criteria should enable detection of the location of the PE segment in each image accurately.

In a second step (step b), a threshold technique combined with an adapted active contour algorithm is applied. It identifies the shape's alterations of the pseudoglottis during vibration. Active contours are widely used to detect object boundaries. Within an image plane, the shape and position of the active contour are governed by internal and external energy terms. The internal energy describes the tension and rigidity of the active contour; the external energy describes the characteristic properties of the image, such as edges, lines, or intensity distribution. The final position of the active contour corresponds to the equilibrium between the internal and external energy terms.

The location of the PE segment in all images of a sequence is assumed to be approximately constant. Drift motions between the patient and the position of the endoscope tip are negligible because PE vibrations are much faster (80-200 Hz) than the unconscious relative movements of the endoscope in a short interval (<100 milliseconds). This fixed location of the PE segment simplifies the second step. A stable tracking of PE deformations is achieved by integrating the knowledge about PE location and structure into the active contour model. By applying these algorithms, the time-dependent deformations of the PE segment can be derived quantitatively and set in relation to the temporal structure of the acoustic signal.

RESULTS

The clinical assessment of the high-speed technique for the endoscopic examination of the substitute voice generator yields the following results.

HIGH-SPEED RECORDINGS

Figure 2 illustrates the wide range of different PE structures and complex oscillation movements. In recording L2, L3, and L5, the PE segment is rather circular; in L4, L6, L7, L8, L9, and L10, the PE segment is more split shaped; and in L1, the PE segment is triangle shaped. However, a quasi-periodic opening and closing of the pseudoglottis can be observed in all recordings. Some sequences show slightly lower quality arising from the examination (ie, illumination) or technical limitations (ie, spatial resolutions). The movements of the PE segment follow a complex quasi-periodic vibration pattern. No chaotic movement occurs even though the vibrations are not completely regular.

IMAGE PROCESSING

The automatic image processing algorithm succeeded in automatic identification of the pseudoglottis in each recording independently from its location, shape, and deformation during oscillation. The tracking algorithm is able to follow the alterations of the shape of the pseudoglottis during vibration even in high-speed recordings with lower image quality. The structure of the PE segment does not significantly affect the segmentation procedure, since we do not use fixed default specifications of PE geometrics within the model. Also the direction of oscillation is well detected. Therefore, even in the recording of L3, where no considerable change of the pseudoglottis area can be observed, the algorithm is able to follow the horizontal shifting of the pseudoglottis.

ACOUSTIC SIGNAL

The acoustic signals corresponding to the high-speed recordings are shown in Figures 3, 4, and 5. The amplitude spectra reflect the quasi-periodic pattern of the tracheoesophageal voices with a fundamental frequency and its spectral harmonics. In general, fundamental frequencies are hard to detect in severely disturbed voices with varying duration of each period, especially in the substitute alaryngeal voice. In this study, it is shown that the fundamental frequency of the substitute voice can be derived by the combined evaluation of data from the high-speed sequences and the acoustic signal. It is defined as the lowest detectable peak of oscillation frequency in the amplitude spectra of both data sets. For example, according to this definition, the fundamental frequency can be found at approximately 190 Hz in L1. The fundamental frequencies of the tracheoesophageal voices L1 through L10 range from 80 to 200 Hz with a different number of harmonics.

ACOUSTIC SIGNAL VS RESULTS OF THE IMAGE PROCESSING

According to the acoustic signal of the sequences, a similar quasi-periodic interval can also be identified in the corresponding pseudoglottis vibration patterns (Figures 3-5). In almost all recordings, they show a strong correlation. The fundamental frequencies of both analyzed pitches can be considered identical within the precision of measurement, which is determined by the sampling rate of approximately 3704 or 4000 Hz and the sequence length of 352 frames. However, L3 does not show high correlation between the PE area and the acoustic signal. Further insight into the recordings showed that the PE segment of L3 in fact moved horizontally without considerable change in the areas of the pseudoglottis. When comparing the oscillating shifting movement to the acoustic signal, the amplitude spectra show similar fundamental frequencies at 178 Hz (Figure 5).

Although the video frame rate would theoretically be able to identify frequencies up to 2000 Hz, no frequencies higher than 750 Hz were detected. This may result from the restricted spatial resolution of the camera, which does not detect slight oscillation amplitudes. Therefore, all demonstrated amplitude spectra are plotted for frequencies up to 750 Hz only.
Figure 2. Segmentation of 10 high-speed recordings of L1 through L10 during 5 frames with an interval between each frame of 1.35 milliseconds (L1-L5 and L9) and 1.25 milliseconds (L6-L8 and L10). Within the sequences, the quasi-periodic opening and closing of the pharyngoesophageal segments can be seen.
Figure 3. Analysis of pseudoglottis vibrations for recordings L1 through L5. A, The areas of the pseudoglottis (a) are demonstrated during an interval that corresponds to 352 video frames. B, Amplitude spectra A of the area of the pseudoglottis. C, Amplitude spectra H of the simultaneously recorded acoustic signal. The areas and amplitude spectra of the pseudoglottis correspond well to each other, except in L3.
Figure 4. Analysis of pseudoglottis vibrations for recordings L6 through L10. For a description, see the legend to Figure 3.
Knowledge about the relationship between PE dynamics and the acoustic signal may help to detect which particular property of the PE segment is responsible for voice quality. Thus, investigating substitute voice production has to analyze both the dynamics of the pseudoglottis and the emitted acoustic signal.

A high-speed technique enables the capture of irregular and very fast movements of the glottis or the pseudoglottis in real time because of its high temporal resolution. Since the high-speed technique focuses just on the detection of fast time-dependent processes, without a loss of information, image recording is performed in black and white instead of color and its processing restricted to gray scales.

The vibrating PE segment structures and movements vary (Figure 2) among different patients who have undergone laryngectomies. Functional properties are therefore hard to describe. An image-processing procedure was developed to identify the PE segment in a high-speed sequence and to track its movements automatically. Using this algorithm, the partially irregular dynamic process of the vibration of the pseudoglottis was analyzed at high temporal resolution. The algorithm succeeded in locating the PE segment (ROI) even in recordings with low image quality. The PE vibrations could be tracked with reasonable accuracy. During sustained phonation, PE vibrations follow a quasi-periodic opening and closing process. In all recordings, the PE cycles and frequencies could be determined even in complex horizontal vibration patterns (eg, in L3). Thus, the image-processing procedure complies with the requirements to be independent of location, shape, and vibration pattern of different PE segments. It reliably detects and tracks the oscillations of the PE segments. Although 3-dimensional effects cannot be differentiated, information about individual oscillation patterns can be derived precisely.

The frequencies of the pseudoglottis oscillations in high-speed recordings resemble the frequencies of the acoustic signal. With good accuracy, the fundamental frequencies and the first harmonics are identical. The differences between the spectral components of the acoustic signal and the PE segment vibrations arise from the limited spatial resolution of the high-speed camera, the limited length of the analyzed high-speed sequence, and modulations of the emitted pseudoglottal acoustic signal by the vocal tract that influence the composition of the acoustic spectra. The agreement of the fundamental frequencies is consistent with videofluoroscopy investigations, which found a strong correlation between morphologic and dynamic properties of the PE segment and substitute voice and proved the PE segment to be the substitute voice generator.

Up to now, it was not possible to indicate the point of the substitute voice’s origin precisely. The image processing now enables us to localize this point exactly. It is shown to be the detected contour of the pseudoglottis.

Further investigation of the relationship between substitute voice quality and PE vibrations demands extended quantitative evaluations of digital high-speed recordings in many patients who have undergone laryngectomies. Metric information on the structure and oscillations of the PE segments could be achieved by using a laser projection system. The analyzed data will be interpreted by applying biomechanical models similar to models of the glottis, the most well-known being the 2-mass model by Ishizaka and Flanagan. The models could virtually demonstrate effects of the PE segment's shape on the functional properties and the acoustic signal. Furthermore, virtual alterations of the PE segment's shape and elasticity of the mucosa could be performed and evaluated. Thus, we would be able to demonstrate how surgical procedures or adapting therapies could influence the PE segment's function to improve and create specific and systematic therapy for voice restoration in patients undergoing laryngectomies.

Submitted for Publication: October 11, 2004; final revision received April 27, 2005; accepted June 22, 2005.

Correspondence: Maria Schuster, MD, Department of Phoniatrics and Pedaudiology, University Hospital, Bohlenplatz 21, 91054 Erlangen, Germany (maria.schuster@phon.i.med.uni-erlangen.de).

Financial Disclosure: None.

Funding/Support: This work was supported by grants from the Deutsche Forschungsgemeinschaft (German Research Council, Bonn, Germany), SFB 603, subproject No. B5.

Figure 5. Horizontal centroid shift of the pharyngoesophageal area (a) during an interval of 95 milliseconds for the recording L3. The dominant frequencies in the amplitude spectrum of the oscillating centroid (A) can also be identified in the amplitude spectrum of the simultaneously recorded acoustic signal (H).
Additional Information: The authors had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

REFERENCES