Published in final edited form as:

J Voice. 2018 March ; 32(2): 256.e1–256.e12. doi:10.1016/j.jvoice.2017.05.014.

Temporal Segmentation for Laryngeal High-Speed Videoendoscopy in Connected Speech

Maryam Naghibolhosseini^a, Dimitar D. Deliyski^{a,b}, Stephanie R.C. Zacharias^{b,c,d,e},
Alessandro de Alarcon^{b,d}, and Robert F. Orlikoff^f

^aDepartment of Communicative Sciences and Disorders, Michigan State University, 1026 Red Cedar Rd, East Lansing, MI, USA

^bDivision of Pediatric Otolaryngology, Cincinnati Children's Hospital Medical Center, 3333 Burnet Ave, Cincinnati, OH, USA

^cDivision of Speech-Language Pathology, Cincinnati Children's Hospital Medical Center, 3333 Burnet Ave, Cincinnati, OH, USA

^dDepartment of Otolaryngology Head and Neck Surgery, University of Cincinnati, 3230 Eden Ave, Cincinnati, OH, USA

^eDepartment of Communication Sciences and Disorders, University Cincinnati, 3202 Eden Ave, Cincinnati, OH, USA

^fCollege of Allied Health Sciences, East Carolina University, 3206 Health Sciences Bldg., Mail Stop 668, Greenville, NC, USA

Abstract

Objective—This study proposes a gradient-based method for temporal segmentation of laryngeal high-speed videoendoscopy (HSV) data obtained during connected speech.

Methods—A custom-developed HSV system coupled with a flexible fiberoptic nasolaryngoscope was used to record one vocally normal female participant during reading of the “Rainbow Passage.” A gradient-based algorithm was developed to generate a motion window. When applied to the HSV data, the motion window acted as a filter tracking the location of the vibrating vocal folds. The glottal area waveform was estimated using a statistical-based image-processing approach. The vocal fold vibratory frequency was computed by an autocorrelation-based extraction of the fundamental frequency (f_0) from the glottal area waveform. Temporal segmentation was then performed based on the f_0 contour and automatic detection of the epiglottic obstructions. Additionally, visual temporal segmentation was performed by viewing the HSV

Address for Correspondence: Maryam Naghibolhosseini, PhD, Department of Communicative Sciences and Disorders, Michigan State University, 1026 Red Cedar Road, Rm 12, East Lansing, MI 48824, Office tel.: (517) 884-8869, naghib@msu.edu.

Portions of this study were presented at the 10th International Conference on Voice Physiology and Biomechanics, Viña del Mar, Chile, March 14–17, 2016, and the 46th Symposium of The Voice Foundation: Care of the Professional Voice, Philadelphia, PA, May 31–June 4, 2017.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

images frame by frame to determine the time points of the vocalization onsets and offsets, and the epiglottic obstructions of the glottis.

Results—The time points resulting from the automatic and visual temporal segmentation methods were cross-validated. The f_0 -contour patterns of rise and fall resulting from the automatic algorithm were found to be in agreement with the visual inspection of the vibratory frequency change in the HSV data.

Conclusions—This study demonstrated the feasibility of automatic temporal segmentation of HSV imaging of connected speech, which allows for mapping the video content into onsets, offsets, and epiglottic obstructions for each vocalization. Automated analysis of HSV imaging of connected speech has significant clinical potential for advancing instrumental voice assessment protocols.

Keywords

Connected Speech; Voice Assessment; High-Speed Videoendoscopy; Laryngeal Imaging

1. Introduction

As communicative disorders, voice problems are revealed in the context of connected speech [1, 2, 3, 4, 5, 6]. However, in traditional clinical practice, voice assessment is mostly performed on vocalizations. That is, connected speech is evaluated either using perceptual assessment or flexible videoendoscopic inspection of the laryngeal function. However, objective instrumental assessment of the vocal fold vibrations via videostroboscopy and acoustic analysis are performed on sustained phonations [7, 8, 9, 10].

Clinical examination of the larynx during voice production is typically performed using videoendoscopy with stroboscopy (videostroboscopy) [7, 11, 12, 13], which does not provide information regarding all aspects of vocal fold dynamics. The use of videostroboscopy is limited due to the lack of ability to assess vocal fold vibrations during connected speech because videostroboscopic images are synchronized with the cycles of the acoustic signal and are produced by sequencing the images at different phases of the glottal cycle [14, 15]. Videostroboscopy can produce apparent slow-motion view of the vocal folds in highly repetitive vibratory cycles, which can be achieved only during sustained phonations [15, 16].

Laryngeal high-speed videoendoscopy (HSV) is an imaging technique for recording the vibrations of the vocal folds with true high-temporal resolution that does not rely on periodic cycles [15, 17, 18, 19]. HSV has been demonstrated as a powerful tool for studying the vocal function and its underlying mechanisms, and has offered clinical improvements over videostroboscopy when used in sustained phonation [14, 20, 21, 22, 23]. Technological constraints historically impeded the implementation of HSV for connected speech and, generally, there is a paucity of research addressing laryngeal vibratory behavior during connected speech.

Recent technological advancements have allowed the successful coupling of HSV systems with flexible fiberoptic nasolaryngoscopes [24, 25]. This provides a unique opportunity to

record vocal fold vibrations during connected speech for the first time. Successful development of effective methods for visualization and automated measurement of the phonatory mechanisms during connected speech can substantially enhance the clinical voice practice. This may afford improvement in assessing vocal fold dysfunctions that are not currently amenable to instrumental voice assessment using laryngeal imaging.

A critical first step in the development of image-processing methodology for HSV analysis of connected speech is to segment the vocalizations over time, a procedure termed “temporal segmentation.” Temporal segmentation is necessary to precede the spatial segmentation, performed to provide analytic representation of the edges of the vocal folds, because the method has to determine where in time the vibrating vocal folds are present before the vocal fold edges can be segmented. Although there is a solid body of published reports in spatial segmentation of the vocal folds, they are not related to the focus of this study. There are no previous studies on temporal segmentation of connected speech. The only reports describing temporal segmentation were designed for sustained phonations and did not provide specific results [14, 18].

The purpose of this study was to develop and validate a method for temporal segmentation of HSV data obtained during connected speech. For this purpose, a custom-built HSV system was used to obtain data from one participant during connected speech. A histogram analysis along with a gradient-based noise removal procedure was performed to extract the vibrating vocal folds from the HSV images and the background noise. An image-processing approach based on moments of inertia was used to determine a motion window that could capture the vibrating vocal folds in each frame of the HSV recording. The motion window was then applied to estimate the glottal area waveform (GAW). The fundamental frequency (f_0) of the GAW was used to estimate the frequency of vibration of the vocal folds and to perform the temporal segmentation.

2. Methods

2.1. HSV Data Acquisition

A vocally normal 38-year-old woman was recruited to participate in this study. The examination was performed at the Center for Pediatric Voice Disorders, Cincinnati Children’s Hospital Medical Center and was approved by the Institutional Review Board of the said hospital. The participant did not have any history of voice disorder. HSV data were recorded during production of the “Rainbow Passage” using a custom-built HSV system at 4,000 frames per second (fps), integration time of 249 μ s and recording length of 29.14 s (total of 116,543 frames). The temporal resolution of 4,000 fps has been previously shown to be clinically acceptable for voice assessment [26]. The spatial resolution of the HSV images was set to 256×256 pixels. The HSV system included a FASTCAM SA-Z color high-speed camera (Photron Inc., San Diego, CA) with 12-bit color image sensor and 64 GB of cache memory, a 300-W xenon light source, model 7152A (PENTAX Medical Company, Montvale, NJ), and a 3.6-mm Olympus ENF-GP Fiber Rhinolaryngoscope (Olympus Corporation, Tokyo, Japan). After recording, the HSV sequence was saved as an uncompressed 24-bit RGB AVI file.

2.2. Visual Temporal Segmentation

Visual temporal segmentation was performed by viewing the HSV images frame by frame. The time points of the vibratory onsets and offsets, as well as the epiglottic obstructions of the glottis were manually determined and recorded as “timestamps.” A highly-trained rater with over 7 years of professional experience in image analysis performed the visual segmentation using Phantom Camera Control Software, version 2.6 (Vision Research Inc., Wayne, NJ). The software allowed the rater to change the video playback speed arbitrarily from 10 to 100 fps (default 30 fps), pause, and advance frame by frame forward and backwards, allowing for optimal determination of the timestamps. The rater repeated the procedure of viewing the vibratory onsets and offsets and the epiglottic obstructions several times to ensure reliable selection of timestamps. These visually segmented timestamps were further used as a reference in the validation of the automatic temporal segmentation.

2.3. Automated Temporal Segmentation

The purpose of the analysis was to extract the vocalized segments and epiglottic obstructions of the glottis from the laryngeal HSV recordings of the “Rainbow Passage.” The analytic procedures included the following steps. A histogram analysis was performed first to remove the saturated pixels and to enhance the contrast of the images. Then, to remove the noise and suppress the motions unrelated to the vocal fold vibrations, a gradient-based algorithm was developed, which was applied to the HSV data through a moving-average procedure. The first and second image moments of inertia after denoising were computed to create a region of interest (ROI), in which the motion of the vocal folds occurred across the frames. The ROI was defined as the union of the areas in which the vibrations of the vocal folds occurred across the frames. The ROI was applied to the denoised frames and the first and the second moments were recomputed to determine a motion window. The appropriate motion window was applied to each frame and acted as a high-pass filter capturing the spatial location of the vibrating vocal folds in each frame of the video sequence. After applying the motion window, the first and second moments of inertia of the filtered images were computed to determine the GAW. Lastly, the f_0 of the vibrations was extracted from the GAW and the temporal segmentation was performed. All the algorithms were coded in 64-bit MATLAB 2015b (MathWorks Inc., Natick, MA).

2.3.1. Histogram Analysis—Color images consist of three separate image components: red, green, and blue, which are termed channels. The pixel intensities for the red channel of one in every 10 frames were sampled for the histogram analysis of the video. The red channel was selected because most of the relevant information is in red, whereas the green and blue channels contain less and mostly redundant information. The histogram analysis was performed to remove the saturated pixels that occur due to light reflections and the dark areas outside the endoscopic aperture. To do this, an upper threshold of 252 was used assuming that pixel intensities larger than this value were associated with the light reflections (RGB images are represented by 8-bit unsigned integer values, ranging from 0 to 255 per channel). A lower threshold of 3 was considered to characterize the dark areas of the images. The median of the intensities of the red pixels was computed after removing the saturated and dark pixels because the relevant information regarding vocal fold vibration is statistically present in intensities between 20% of the median and the median. Then a linear

contrast enhancement was performed by normalizing the intensity values in the aforementioned range.

2.3.2. Noise Removal—After the histogram analysis of the data, a gradient-based algorithm was developed to remove the excessive noise from the HSV images and to track the location of the vibrating vocal folds over time. In this noise-removal procedure, a gradient-based algorithm using a moving-average filtering was applied. The size of the moving average window was chosen to be 50 ms with a step size of 10 ms, because the relative motion between the endoscope and the tissues is of low frequency (i.e., less than 15 Hz). The gradient-based algorithm that was performed for each window of the moving average is explained, as follows.

The gradient was found by computing the time differential with delta of 1.5 ms (6th frames). A Gaussian smoothing (blur) with standard deviation of 1 was applied spatially to the quadratic value of the gradient to remove part of the noise and enhance the glottal vibratory patterns. The result of Gaussian smoothing was then filtered temporally using a 15th-order Hamming bandpass filter with cutoff frequencies of 70 and 1,000 Hz. Such filtering was applied to remove the low- and high-frequency noise outside the frequency range of vocal fold vibration. The resulting image sequence is further termed spatio-temporally filtered gradient.

A challenge that we encountered in the above gradient-based algorithm was the presence of edges in the gradient that were not associated with the vocal fold vibrations but rather with the movements of the epiglottis or other laryngeal tissues. To remove these edges, the gradient was smoothed, first temporally, and then spatially using a Gaussian filter. The absolute value of the smoothed gradient was then added to itself (termed positive gradient) to remove the outward edge movements (e.g., anterior-to-posterior motion of epiglottis). The absolute value was also subtracted from itself (termed negative gradient) to remove the inward edge movements. Therefore, the result was two sets of image frames; in one of each set, the edge movements in one direction were omitted. To further reduce the noise, the positive gradient and the negative gradient were bandpass-filtered with cutoff frequencies of 70 and 1,000 Hz. The absolute values of the filtered positive and negative gradients were then multiplied by the spatio-temporally filtered gradient, found in the first step, to remove the edges appearing in the gradients due to motions not associated with the vocal fold vibrations. The resulting denoised gradient was used to design the motion window.

2.3.3. Motion Window—The motion window was designed based on the first and second moments of the denoised gradient (see [27] for complete description of the moments). The first moment, M_1 , determines the centroid of the denoised gradient, which corresponds to the center of the glottis. The second moment, M_2 , corresponds to the area of the glottis. The first and second horizontal moments, denoted by $M_1(x, t_i)$ and $M_2(x, t_i)$, and the first and second vertical moments, denoted by $M_1(y, t_i)$ and $M_2(y, t_i)$, are computed as follows:

$$\begin{aligned}
M_1(x, t_i) &= \sum_{x=1}^{256} \sum_{y=1}^{256} f(x, y, t_i) x / \sum_{x=1}^{256} \sum_{y=1}^{256} f(x, y, t_i) \\
M_1(y, t_i) &= \sum_{x=1}^{256} \sum_{y=1}^{256} f(x, y, t_i) y / \sum_{x=1}^{256} \sum_{y=1}^{256} f(x, y, t_i) \\
M_2(x, t_i) &= \sqrt{\sum_{x=1}^{256} \sum_{y=1}^{256} f(x, y, t_i) x^2 / \sum_{x=1}^{256} \sum_{y=1}^{256} f(x, y, t_i) - M_1(x, t_i)^2} \\
M_2(y, t_i) &= \sqrt{\sum_{x=1}^{256} \sum_{y=1}^{256} f(x, y, t_i) y^2 / \sum_{x=1}^{256} \sum_{y=1}^{256} f(x, y, t_i) - M_1(y, t_i)^2}
\end{aligned}$$

Based on the first and second moments, the ROI was found. The edges of the ROI were smoothed by filtering using a 7th-order Hamming filter. The ROI was multiplied by the denoised gradient, which resulted in the removal of noticeable amount of noise. The first and second moments were recomputed after the noise removal. A spline interpolation of the first moment was computed next to estimate the first moments for all the frames. This was performed because the denoised gradient was found for the middle frame of each window of the moving average data. The motion window was then designed as an ellipse, with the center coordinates being the first horizontal and vertical moments and the diameters being the maxima of the horizontal and vertical second moments. In fact, the motion window was a 256×256 image that had the values of 1 inside the aforementioned ellipse and values of 0 outside it. The edges of the motion window were smoothed using a 7th-order Hamming filter. Applying such motion window to the original HSV image frames resulted in the preservation of the areas associated with the vocal folds motion containing the glottis, while excluding the rest in the images, which contained light reflections, glottis reflections, and supraglottal tissues.

2.3.4. Glottal Area Waveform—The GAW was estimated by computing the horizontal and vertical second moments of the images of the vibrating vocal folds after applying the motion window to each inverted HSV frame. The HSV images were inverted because the second moments provide the width and height of the brightness; therefore, to find the width and height of the darkness (i.e., glottis), the images had to be inverted.

The root mean square (RMS) of the horizontal and vertical second moments represented the area change of the glottis and was used to find the GAW. The RMS of the second moments was upsampled to 50 kHz using spline interpolation. This waveform was then high-pass filtered with an approximate cutoff frequency of 50 Hz to remove the low-frequency noise, resulting in an estimate of the GAW.

2.3.5. Fundamental Frequency Extraction—The f_0 -extraction algorithm was based on autocorrelation. The GAW was first windowed with a window size of 40 ms and a window advance of 2.5 ms. The autocorrelation of GAW was computed for each window where the peak value in the delay range of 2.5–15 ms (corresponding to 66.6–400 Hz) was determined. If the autocorrelation peak value was larger than 70% of the total window energy, and the total window energy was exceeding 5% of the maximum of the sample energy contour, that window was classified as voiced. Finally, the f_0 was computed as the inverse value of the delay corresponding to the autocorrelation peak in the voiced window.

2.3.6. Epiglottic Obstruction Detection—Epiglottic obstructions were detected though analyzing the pixel intensities of the red channel in the HSV images. The sum of intensities

of the red channel for each individual frame was computed. Preliminary investigation showed that when the epiglottis moves in the anterior-to-posterior direction toward the glottis, a large reflection occurs along the epiglottis, which saturated a substantial part of the image once it reached the point of obstruction of the glottis. This saturation added to the replacement of the dark pixels corresponding to the glottis, with brighter epiglottis pixels resulting in large intensity values in the corresponding frames. Therefore, the automated threshold for detecting epiglottic obstructions of the glottis was set at the 90th percentile of the image intensity. That is, when the integral of the intensity values of the red channel of a frame exceeded the 90th percentile of the integral of the HSV sample, the frame was classified as containing epiglottic obstruction.

2.3.7. Automatic Temporal Classification—The final stage of the automatic temporal segmentation was performed using the f_0 contour and the epiglottic obstruction information with the purpose of classifying the continuously voiced segments by separating them from outliers. Consecutive voiced windows with durations larger than 30 ms were classified as voiced segments. Additionally, segments with epiglottic obstructions occurring during the onset or offset of vibrations with total voiced duration greater than 30 ms were classified as voiced segments. Pauses in the f_0 contour of phonatory segments with durations shorter than 5 ms were also considered as part of the voiced segments, because of the finding that they were typically corresponding to irregular vibrations during phonatory laryngeal adjustments. Finally, the epiglottic obstructions were excluded from the voiced segments and classified as unvoiced because vocal fold vibrations are not visible, thus preventing further analyses.

3. Results

The histogram of the red pixels is shown in Fig. 1. The pixels with intensities between 20% of the median and the median were anticipated to include the information regarding the vocal folds motion and were used for further analysis of the HSV data; the boundaries of this region is marked by the two stars in the figure. For our participant, the median was 42 and 20% of the median was 8.4. The histogram was then stretched between these two values (i.e., 8.4 and 42) for the goal of contrast enhancement.

After the histogram enhancement, the moving-average gradient-based algorithm was applied to the HSV data. An example of the gradient for the red channel is shown in Fig. 2-a. The area in the center with the largest intensity corresponds to the vibrating vocal folds. The circular area around the main peak and the nonzero values corresponds to the aperture of the endoscope and the non-zero intensities within the circle are noise. In the quadratic gradient shown in Fig. 2-b, the area associated with the vibrating vocal folds has much higher intensity in comparison with the noise. After Gaussian smoothing of the quadratic gradient, most of the noise is removed as shown in Fig. 2-c.

To remove the edges from the gradient, the positive and negative gradients were found after smoothing the gradient temporally and spatially. Examples of the positive and negative gradients for two particular time points of interest are shown in Fig. 3. In Fig. 3-a and 3-b, the epiglottis is moving in the anterior-to-posterior direction between the two frames of the gradient, thus an edge appears in the negative gradient (Fig. 3-b) and no edge is present in

the positive gradient (Fig. 3-a). In Fig. 3-c and 3-d, the epiglottis was moving in the posterior-to-anterior direction, thus the edge is formed only in the positive gradient (Fig. 3-c).

Following band-pass filtering, the absolute values of the positive and negative gradients were multiplied. Two examples of the average of 200 frames after such multiplication are shown in Fig. 4-a and 4-c. The filtered gradients were then multiplied by the spatio-temporally filtered gradient and averaged, which were termed denoised gradients, shown in Fig. 4-b and 4-d for the two frame sets. As seen in Fig. 4-b and 4-d, the noise was reduced noticeably when the data were multiplied with the spatio-temporally filtered gradient.

The ROI, which was found from the first and second moments of the denoised gradient, is shown in Fig. 5. The two- and three-dimensional views of the ROI are shown in Fig. 5-a and 5-b, respectively. The ROI values are 1 within the elliptical area and 0 outside this area. The transition from 0 to 1 is smooth due to filtering.

After applying the ROI to the denoised gradient, the first and second moments were recomputed. Applying the ROI to the denoised gradient significantly improved the horizontal and vertical first-moment estimates of the coordinates of the glottis, because the effects of the variations due to noise were suppressed. The diameters of the motion window were estimated using the values of the horizontal and vertical second moments. For the HSV sample used in this study, the diameters were found to be 34 and 23 pixels. The two- and three-dimensional views of the motion window are shown in Fig. 6-a and 6-b, respectively. The center of the motion window changed along the frames depending on the values of the first moments to follow the location of the vocal folds.

The appropriate motion window was applied to each HSV frame to capture the vibrating vocal folds. Examples of applying the motion window to two arbitrary frames are shown in Fig. 7. The original HSV frames are shown in Fig. 7-a and 7-c, and the results of applying the motion window to the HSV frames are shown in Fig. 7-b and 7-d, respectively. Fig. 7 demonstrates how successful the motion window was in capturing the vocal folds in these frames, both for their location and size. A visual inspection of the entire HSV recording length of 29.14 s (total of 116,543 frames) confirmed that the motion-window algorithm did not incur any errors.

The high-pass filtering of the upsampled RMS second moment waveform excluded the low-frequency motions and preserved the areas associated with the vocal folds vibrations, which corresponded to the GAW. The estimated GAW for the HSV recording of the entire “Rainbow Passage” is shown in Fig. 8-a; Fig. 8-b shows the GAW zoomed-in between 1–1.5 s.

The resulting f_0 contour of the vocal fold vibration for the entire “Rainbow Passage,” which was computed from the GAW using the autocorrelation technique, is shown in Fig. 9-a. Fig. 9-b shows the f_0 contour zoomed-in between 1–1.5 s.

Video S1 in the Appendix illustrates some of the key stages of the temporal segmentation method. The upper left quadrant shows an HSV sequence in the range from frame number

4,000 to 5,900. The result of applying the motion window to the HSV frames is playing synchronously on the bottom left. The corresponding GAW and f_0 contour between frame numbers 4,000–5,900 are shown on the right. The dark dot moving along the GAW and f_0 contour corresponds to the HSV frame playing on the video. As can be seen on the bottom left, the ellipse-shaped motion window follows the location of the vibrating vocal folds across the frames, once vibration begins. The f_0 values for the first few cycles at the beginning of the vibration are zero due to the aperiodic nature of the phonatory onset.

The automatically computed f_0 for the entire “Rainbow Passage” along with the visual and automatic temporal segmentation of voicing and epiglottic obstructions are shown in Fig. 10. The segmentation results are shown along with the corresponding sentences of the “Rainbow Passage.” The voiced segments that were determined using the visual and automatic temporal segmentations are shown by the black and blue horizontal lines, respectively. The segments with epiglottic obstruction detected by the visual and automatic methods are shown by the green solid lines and the highlighted green areas, respectively. Although good agreement can be seen in Fig. 10, there are some discrepancies between the automatic and visual temporal segmentation results (e.g., around 4 s).

In addition to the agreement between the automatic and visual segmentation, the f_0 -contour patterns of rise and fall resulting from the automatic algorithm were in agreement with the visual inspection of the vibratory frequency changes in the HSV video sequence.

4. Discussion

Based on the results, the gradient-based algorithm using a moving-average procedure was successfully developed to remove the noise and information unrelated to vocal fold vibrations and to enhance the vibrating vocal folds in the images. The results also demonstrate the success of this algorithm in removing the edges that appeared in the gradient but were not related to the vocal fold vibrations, such as the edges created by the movements of the epiglottis. The small remaining noise in the images was removed successfully by applying the ROI to the images. The comparison between the first moment before and after applying the ROI showed noticeable differences due to noise removal.

The implemented motion window was able to follow the vibrating vocal folds across the frames and remove the information in the frames that were not associated with the vibrating vocal folds. The visual inspection of the data showed that the motion window deviated from the actual location of vocal folds only in frames without vibrations of the vocal folds, which was not an issue because the goal was to segment the vibrating vocal folds. The GAW was found successfully using the second moment of each frame post applying the motion window. The f_0 -contour patterns of rise and fall were in agreement with the visual inspection of the vibratory frequency change in the HSV data.

Temporal segmentation was performed automatically based on the f_0 contour and the epiglottic obstructions data, then cross-validated by comparing the results to the visually segmented onsets and offsets of vocalizations and the epiglottic obstructions of the glottis. The main reason for the observed minor discrepancies were the occasional aperiodic vocal

fold vibrations, which were not determined as voiced either during pitch extraction or automatic segmentation due to the required minimum onset-to-offset duration of 30 ms. In some instances the discrepancies were caused by inaccuracies in the visual segmentation due to poor image quality or inability to judge the exact beginning of oscillation. The detection of the epiglottic obstructions using the automatic and visual methods was in excellent agreement.

5. Conclusions

This study used for the first time a custom-built HSV system to produce recordings during connected speech. Using HSV to record endoscopic images of the vocal folds during vocalizations in connected speech is a major breakthrough which can lead to significant improvements in instrumental voice assessment and clinical management of voice disorders. However, the development of automated image processing methodology to aid the analysis and the interpretation of the HSV recordings of connected speech is imperative, because it is not feasible to visually assess the data practically in the clinic. This work is an essential step toward measuring meaningful vibratory characteristics to explain the mechanics of vocal fold vibrations. The performance of the algorithm was shown to be satisfactory by comparison of the visually segmented voice onsets and offsets of vocalizations and epiglottic obstructions. The developed automatic algorithm for temporal segmentation was able to analyze the HSV data for the entire “Rainbow Passage” in a few minutes and provided accurate and reliable information regarding the onsets and offsets of vocalizations, and the epiglottic obstructions of the glottis. Although the results are very encouraging, a larger sample from more participants is necessary to address likely intra- and inter-subject reliability issues and to optimize the algorithm for various tokens of connected speech.

Temporal segmentation is a necessary step toward the development of a new-generation automated measurement methods for laryngeal imaging. The ultimate goal is to develop new clinical protocols for instrumental voice assessment of connected speech.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding for this study was provided by the National Institutes of Health, NIDCD, R01 DC007640 “Efficacy of Laryngeal High-Speed Videoendoscopy” and by the Michigan State University Foundation through the MSU Foundation Professor.

References

1. Morrison MD, Rammage LA. Muscle misuse voice disorders: description and classification. *Acta oto-laryngologica*. 1993; 113(3):428–434. [PubMed: 8517149]
2. Yiu E, Worrall L, Longland J, Mitchell C. Analysing vocal quality of connected speech using Kay’s computerized speech lab: a preliminary finding. *Clin Linguist & Phon*. 2000; 14(4):295–305.
3. Halberstam B. Acoustic and perceptual parameters relating to connected speech are more reliable measures of hoarseness than parameters relating to sustained vowels. *ORL*. 2004; 66(2):70–73. [PubMed: 15162004]

4. Roy N, Gouse M, Mauszycki SC, Merrill RM, Smith ME. Task specificity in adductor spasmodic dysphonia versus muscle tension dysphonia. *The Laryngoscope*. 2005; 115(2):311–316. [PubMed: 15689757]
5. Maryn Y, Corthals P, Van Cauwenberge P, Roy N, De Bodt M. Toward improved ecological validity in the acoustic measurement of overall voice quality: combining continuous speech and sustained vowels. *J Voice*. 2010; 24(5):540–555. [PubMed: 19883993]
6. Lowell SY. The acoustic assessment of voice in continuous speech. *SIG 3 Perspectives on Voice and Voice Disorders*. 2012; 22(2):57–63.
7. Stemple, JC., Glaze, LE., Gerdeman, BK. *Clinical voice pathology: Theory and management*, Cengage Learning. 2000.
8. Stojadinovic A, Shaha AR, Orlikoff RF, Nissan A, Kornak MF, Singh B, Boyle JO, Shah JP, Brennan MF, Kraus DH. Prospective functional voice assessment in patients undergoing thyroid surgery. *Ann Surg*. 2002; 236(6):823–832. [PubMed: 12454521]
9. Mehta DD, Hillman RE. Voice assessment: updates on perceptual, acoustic, aerodynamic, and endoscopic imaging methods. *Curr Opin Otolaryngol Head Neck Surg*. 2008; 16(3):211–215. [PubMed: 18475073]
10. Aronson, AE., Bless, D. *Clinical voice disorders*, Thieme. 2011.
11. Kitzing P. Stroboscopy—a pertinent laryngological examination. *J Otolaryngol*. 1985; 14(3):151–157. [PubMed: 4068109]
12. Bless D, Hirano M, Feder R. Videostroboscopic evaluation of the larynx. *J Ear Nose Throat*. 1987; 66(7):289–296.
13. Woo P, Casper J, Colton R, Brewer D. Aerodynamic and stroboscopic findings before and after microlaryngeal phonosurgery. *J Voice*. 1994; 8(2):186–194. [PubMed: 8061775]
14. Deliyski DD, Petrushev PP, Bonilha HS, Gerlach TT, Martin-Harris B, Hillman RE. Clinical implementation of laryngeal highspeed videoendoscopy: challenges and evolution. *Folia Phoniatr et Logop*. 2008; 60(1):33–44.
15. Deliyski DD, Hillman RE. State of the art laryngeal imaging: research and clinical implications. *Curr Opin Otolaryngol Head Neck Surg*. 2010; 18(3):147–152. [PubMed: 20463479]
16. Mehta DD, Hillman RE. Current role of stroboscopy in laryngeal imaging. *Curr Opin Otolaryngol Head Neck Surg*. 2012; 20(6):429–436. [PubMed: 22931908]
17. Hertegard S, Lindestad P. Vocal fold vibrations studied during phonation with high-speed video imaging, Karolinska Institute. Huddinge University Hospital Phoniatic and Logopedic Progress Report. 1994; 9:33–40.
18. Deliyski, D. Laryngeal high-speed videoendoscopy. In: Kendall, KA., Leonard, R., editors. *Laryngeal evaluation: indirect laryngoscopy to high-speed digital imaging*. Thieme Medical Publishers; 2010. p. 243-270.
19. Patel RR, Dubrovskiy D, Döllinger M. Measurement of glottal cycle characteristics between children and adults: physiological variations. *J Voice*. 2014; 28(4):476–486. [PubMed: 24629646]
20. Patel R, Dailey S, Bless D. Comparison of high-speed digital imaging with stroboscopy for laryngeal imaging of glottal disorders. *Ann Otol Rhinol Laryngol*. 2008; 117(6):413–424. [PubMed: 18646437]
21. Zacharias SRC, Myer CM, Meinzen-Derr J, Kelchner L, Deliyski DD, de Alarcón A. Comparison of Videostroboscopy and High-speed Videoendoscopy in Evaluation of Supraglottic Phonation. *Ann Otol Rhinol Laryngol*. 2016; 125(10):829–837. [PubMed: 27411363]
22. Zacharias SRC, Deliyski DD, Gerlach TT. Utility of Laryngeal High-Speed Videoendoscopy in Clinical Voice Assessment. *J Voice*. 2017 Jun 5. Epub ahead of print.
23. Powell M, Deliyski DD, Hillman R, Zeitels S, Burns J, Mehta D. Comparison of Videostroboscopy to Stroboscopy Derived From High-Speed Videoendoscopy for Evaluating Patients With Vocal Fold Mass Lesions. *Am J Speech-Lang Pathol*. 2016; 25(4):576–589. [PubMed: 27716854]
24. Zañartu M, Mehta DD, Ho JC, Wodicka GR, Hillman RE. Observation and analysis of in vivo vocal fold tissue instabilities produced by nonlinear source-filter coupling: A case study. *J Acoust Soc of Am*. 2011; 129(1):326–339. [PubMed: 21303014]
25. Mehta, DD., Deliyski, DD., Zeitels, SM., Zañartu, M., Hillman, RE. Integration of transnasal fiberoptic high-speed videoendoscopy with time-synchronized recordings of vocal function, in

Normal & Abnormal Vocal Folds Kinematics: High Speed Digital Phonoscopy (HSDP), Optical Coherence Tomography (OCT) & Narrow Band Imaging (NBI®), Volume I: Technology. Vol. 12. San Francisco, CA: Pacific Voice & Speech Foundation; 2015. p. 105-114.

26. Deliyski DD, Powell ME, Zacharias SR, Gerlach TT, de Alarcon A. Experimental investigation on minimum frame rate requirements of high-speed videoendoscopy for clinical voice assessment. Biomed Signal Process Control. 2015; 17:21–28. [PubMed: 28989342]
27. Deliyski DD, Petrushev P. Methods for objective assessment of high-speed videoendoscopy. Proc Adv in Quant Laryngol (AQL). 2003:1–16.

Biographies

Maryam Naghibolhosseini, PhD, is a Postdoctoral Research Associate in the Voice and Speech Lab at the Department of Communicative Sciences and Disorders, Michigan State University (MSU). She received her Bachelor's and Master's degree in Biomedical and Electrical Engineering and completed her PhD in Speech and Hearing Sciences at the Graduate Center of the City University of New York. She is interested in laryngeal imaging, voice biomechanics and modeling, and clinical voice assessment. She aims to apply her engineering skills along with her scientific background to develop new methodologies for voice assessment using laryngeal high-speed videoendoscopy in connected speech.

Dimitar D. Deliyski, PhD, is Professor and Chairperson of the Department of Communicative Sciences and Disorders at MSU. His research interests have focused on improving the clinical assessment of voice and speech disorders, refining the understanding of the mechanisms of voice production, creating new methods for computer imaging of the vocal folds and for acoustic analysis of voice and speech. Recently, he has been selected to receive the 2017 Quintana Research Award at the 46th Annual Symposium of Voice Foundation for his lifetime achievements in designing and utilizing technology and applying engineering principles to voice science.

Stephanie R.C. Zacharias, PhD, CCC-SLP, is an Assistant Professor in the Departments of Communication Sciences and Disorders and Otolaryngology–Head and Neck Surgery at the University of Cincinnati and a Speech-Language Pathologist in the Divisions of Pediatric Otolaryngology and Speech-Language Pathology at Cincinnati Children’s Hospital Medical Center. Clinical and research interests include assessment, treatment, and voice outcomes in pediatric voice. Currently she is working on modelling effects of voice therapy techniques in children, efficacy of high-speed videoendoscopy, assessing physiologic and physical voice changes in boy singers, and assessing different treatment models for improving delivery of pediatric voice care.

Alessandro de Alarcon, MD, MPH, is the Director of the Center for Pediatric Voice Disorders at Cincinnati Children’s Hospital Medical Center, and Associate Professor at the Department of Otolaryngology–Head and Neck Surgery, University of Cincinnati College of Medicine. Since 2008, his attention has been focused on developing new procedures for the management of airway and voice problems. His main interest is in pediatric voice. Specifically, he has been focused on areas that have included common voice problems in children (pediatric vocal fold lesions) as well as more dedication to measurement and evaluation of dysphonia associated with pediatric airway reconstruction.

Robert F. Orlikoff, PhD, CCC-SLP, is Professor and Dean of the College of Allied Health Sciences at the East Carolina University in Greenville, NC. He served previously as the

Associate Dean for Academic Affairs in the College of Education and Human Services, Chair of the Department of Communication Sciences and Disorders, and Director of the Clinical Voice Laboratory at West Virginia University. His research has mainly centered on the acquisition of valid and reliable acoustic, physiologic, and aerodynamic data so that they may provide a firm and objective foundation for the accurate diagnosis and effective treatment of voice disorders.

Figure 1. Histogram of the red channel pixels after saturated pixel removal. The median and 20% of the median are shown by the stars.

Figure 2.

Panel (a): Gradient of the red channel intensities between frames 5,123 and 5,129. Panel (b): Quadratic gradient of panel (a). Panel (c): The result of Gaussian smoothing of the quadratic gradient.

Figure 3.

Panel (a): Positive gradient between frames 12,481 and 12,487. Panel (b): Negative gradient between frames 12,481 and 12,487. Panel (c): Positive gradient between frames 13,281 and 13,287. Panel (d): Negative gradient between frames 13,281 and 13,287.

Figure 4.

Panel (a): Average of the multiplication of filtered gradients between frames 12,481 and 12,682. Panel (b): Averaged denoised gradient between frames 12,481 and 12,682. Panel (c): Average of the multiplication of filtered gradients between frames 13,281 and 13,482. Panel (d): Denoised gradient between frames 13,281 and 13,482.

Figure 5.
Panel (a): Two-dimensional view of the ROI. Panel (b): Three-dimensional view of the ROI, region of interest.

Figure 6.

Panel (a): Two-dimensional view of the motion window. Panel (b): Three-dimensional view of the motion window.

Figure 7. Images for frame 12,481 (panel (a)) and frame 13,281 (panel (c)). The motion window applied to frame 12,481 (panel (b)) and 13,281 (panel (d)).

Figure 8.
Panel (a): The GAW for the entire “Rainbow Passage”. Panel (b): The GAW zoomed-in between 1–1.5 s. GAW, glottal area waveform.

Figure 9.

Panel (a): The Fo-contour of the vocal fold vibrations for the entire “Rainbow Passage”.

Panel (b): The Fo-contour zoomed-in between 1–1.5 s.

Figure 10.

Comparison of the automatically computed f_0 contour (red line) for the entire “Rainbow Passage” along with the visual and automatic temporal segmentation of voicing timestamps, shown by the black and blue horizontal lines, respectively. The green lines show the visual epiglottic obstruction detection results, and the shaded green areas correspond to the automatic algorithm. Each panel shows the result for a different sentence of the “Rainbow Passage”, written above each panel.