Original Paper

Folia Phoniatrica et Logopaedica

Folia Phoniatr Logop 2008;60:33–44
DOI: 10.1159/000111802

Published online: November 30, 2007

Clinical Implementation of Laryngeal High-Speed Videoendoscopy: Challenges and Evolution

Dimitar D. Deliyski[a] Pencho P. Petrushev[b] Heather Shaw Bonilha[a]
Terri Treman Gerlach[c] Bonnie Martin-Harris[d] Robert E. Hillman[e]

[a]Communication Sciences and Disorders and [b]Department of Mathematics, University of South Carolina, Columbia, S.C., [c]Voice and Swallowing Center, Charlotte Eye Ear Nose and Throat Associates, Charlotte, N.C., [d]Evelyn Trammell Institute for Voice and Swallowing, Medical University of South Carolina, Charleston, S.C., and [e]Center for Laryngeal Surgery and Voice Rehabilitation, Massachusetts General Hospital, Boston, Mass., USA

Key Words
Voice assessment · High-speed videoendoscopy · Stroboscopy · Kymography · Vocal fold vibration

Abstract
High-speed videoendoscopy (HSV) captures the true intracycle vibratory behavior of the vocal folds, which allows for overcoming the limitations of videostroboscopy for more accurate objective quantification methods. However, the commercial HSV systems have not gained widespread clinical adoption because of remaining technical and methodological limitations and an associated lack of information regarding the validity, practicality, and clinical relevance of HSV. The purpose of this article is to summarize the practical, technological and methodological challenges we have faced, to delineate the advances we have made, and to share our current vision of the necessary steps towards developing HSV into a robust tool. This tool will provide further insights into the biomechanics of laryngeal sound production, as well as enable more accurate functional assessment of the pathophysiology of voice disorders leading to refinements in the diagnosis and management of vocal fold pathology. The original contributions of this paper are the descriptions of our color high-resolution HSV integration, the methods for facilitative playback and HSV dynamic segmentation, and the ongoing efforts for implementing HSV in phonomicrosurgery, as well as the analysis of the challenges and prospects for the clinical implementation of HSV, additionally supported by references to previously reported data.

Presented as a keynote lecture at the 7th International Conference: Advances in Quantitative Laryngology, Voice and Speech Research AQL, Groningen, The Netherlands, October 2006.

Challenges to Laryngeal High-Speed Videoendoscopy

The most common clinical technique for laryngeal visualization is the videoendoscopy with stroboscopy. The benefit of stroboscopy has been clinically demonstrated [1]. Even though stroboscopy has greatly improved functional evaluation and management of voice disorders, it has inherent and well-known limitations. The intracycle

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

1021–7762/08/0601–0033$24.50/0

Accessible online at:
www.karger.com/fpl

Dimitar Deliyski, PhD
University of South Carolina, Communication Sciences and Disorders
1621 Greene Street, 6th floor
Columbia, SC 29208 (USA)
Tel. +1 803 777 2245, Fax +1 803 777 3081, E-Mail ddeliyski@sc.edu

vibration seen when viewing stroboscopy displays an illusory 'slow motion', produced by sequencing different phases of the glottal cycle obtained across many periods (fig. 1). Therefore, stroboscopy does not provide a true view of the vocal fold vibration. For each video frame, stroboscopy relies on pitch tracking through a laryngeal contact microphone in order to predict the phase of the upcoming glottal cycles, assuming that the glottal period will not change within the next 30-ms video frame. Aperiodic vibration causes the strobe light to become asynchronized with the actual phase of vocal fold movements preventing the visualization in 'slow motion'. As a result, stroboscopy cannot be used on persons whose voice disorder has caused their vocal fold movement to become aperiodic. Thus, many patients cannot benefit from the technology of stroboscopy even though it is considered the current 'gold standard' for laryngeal imaging. Another restriction to the use of stroboscopy is the difficulty in classifying functional voice disorders when it is the sole assessment technique used [2]. The third restriction limits scientific and diagnostic knowledge of the onset and offset of vibration. Onset of phonation has been found to be useful in characterizing types of vocal fold function [3].

Complementing stroboscopy with videokymography (VKG) allows for viewing of the true vibratory features of the vocal folds in the temporal domain for a selected single line across the glottis [4, 5]. The relatively low cost of VKG and the great amount of validation work done by its inventors [4, 5] made it clinically feasible. The main concerns with VKG have been the difficulty in reliably scanning the same location on the anatomical structure due to endoscopic motion and the lack of a full laryngeal view throughout the recording, as well as the limitation of the duration for continuous images to 18 ms. These limitations are being addressed by developing a new generation digital kymography (DKG) technique [6].

Ultimately, high-speed videoendoscopy (HSV) is the only technique that captures the true intracycle vibratory behavior through a full image of the vocal folds. Therefore, HSV by default overcomes the above limitations of stroboscopy and VKG providing for the possibility of more accurate objective quantification of the vocal fold vibratory behavior. Moreover, the existing duality between HSV and DKG [7, 8] makes HSV uniquely suited for either spatial or kymographic representation of the same content. It is important to note that high-speed motion films have been used for studying the motion of the vocal folds [9] long before the development of commercial systems for laryngeal videostroboscopy. The invention

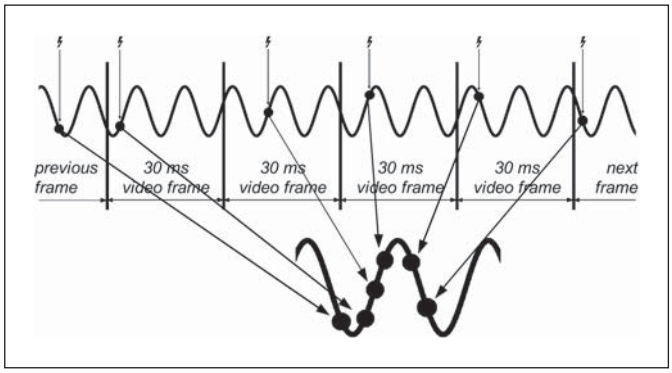

Fig. 1. Principles of videostroboscopy: the strobe light flashes once during each video frame in a moment determined through the phase of a contact microphone signal (top). The recorded images from each frame are then sequenced to create an artificial cycle (bottom).

of videostroboscopy was a technological breakthrough shortcutting the long cycle needed for the technology to meet the demand. Eventually, HSV systems were introduced commercially as recently as the 1990s.

The commercial HSV systems, however, have not gained widespread clinical adoption because of remaining technical, methodological and practical limitations, and an associated lack of information regarding the validity and clinical relevance of HSV. These limitations are highly interrelated, thus, hard to overcome one at a time, making HSV the 'Gordian knot' of laryngeal visualization. That is, the technical limitations can be identified only with clinical experience and can be resolved by the developers when the manufacturers register an increased demand for HSV. But there are no established clinical protocols for HSV, therefore, there is a lack of uniform data for demonstrating its clinical relevance and for guiding the development of the appropriate methodology. Nevertheless, there is no sufficient incentive among clinicians to implement new costly technology with unestablished requirements and unclear benefits, even though it is intellectually understood that it is superior to stroboscopy in many ways. This complex dilemma can be resolved only by addressing all of the above aspects simultaneously in a highly coordinated manner.

Practical Challenges

High cost is the factor most commonly thought to limit the practicality of HSV as a clinical tool. Which cost is that? The cost of HSV equipment has dropped and it is no longer excessively higher than that of stroboscopy. In

fact, stroboscopic equipment was also costly at the time it was first introduced. More precisely, the real limiter of HSV is the unjustified cost. That is, HSV is considered a technique supplementing stroboscopy with its high temporal resolution allowing for the registration of the true intracycle vibratory behavior of the vocal folds. Thus, HSV is regarded as a secondary, supplemental technique, which does not have developed guidelines for clinical use, yet its cost is higher than that of stroboscopy.

In addition, there is no solid evidence yet whether the addition of HSV to the clinical protocol significantly improves the functional assessment of the pathophysiology of voice disorders and/or the diagnosis of laryngeal pathology. Consequently, the health insurance companies do not provide coverage for this additional procedure. Therefore, the cost of the time spent for training, performing the procedure and maintenance is added to the equipment cost, yet there is not a billing/coding procedure in place for financial reimbursement. VKG provided an example of a very useful, low-cost technique that was highly supplemental to and compatible with stroboscopy, which has not become widely used in the 10 years of commercial availability, mainly due to these factors.

Finally, more unanswered questions affect the practicality of HSV: How to store the huge amount of data? How to quickly view the lengthy HSV recordings? Does the bright xenon light pose any risks to the patients? Are the current examination rooms large enough to host two endoscopy systems?

Therefore, the future of HSV as a clinical technique is not in supplementing stroboscopy, but rather in supplanting it. That is, a clinically successful HSV approach has to demonstrate new clinically useful features in addition to possessing all of the useful features of stroboscopy. Such stroboscopy features that need to be included in HSV systems are: real-time visualization of 'slow motion', synchronous audio playback, high-quality color images, high pixel resolution, sharp glottal edges during vibration, and lengthy recordings. All of these features can be realized through HSV as shown further.

Technical Challenges

The strength of HSV is the high temporal resolution allowing for functional imaging of the vocal folds. However, HSV is traditionally perceived as a tool capturing monochromatic images with low contrast, dynamic range, spatial resolution, short sample durations, and narrow viewing angle, inadequate for structural imaging. In 2006, Richard Wolf GmbH (Knittlingen, Germany) introduced the first commercial color HSV system, with temporal resolution up to 4,000 fps, spatial resolution of 256 × 256 pixels, and sample duration of 4 s. In parallel, the University of South Carolina (USC) Voice and Speech Lab integrated a color high-resolution HSV system presented further in this paper. What is the specific importance of the technical factors in the clinic?

Color and spatial resolution are very important factors when identifying lesions, vascularities and tissue changes, and for accurately representing the glottal edges. Spatial resolution is also important since it allows for a wide view angle necessary to examine the full anterior-posterior view of the vocal folds and their surrounding anatomic structures. The temporal resolution is essential for accurately displaying the mucosal wave and providing sharp glottal edges, especially when viewing high-pitched samples. Long sample duration is necessary to register multiple phonations in a continuous recording (i.e. comfortable, high and low pitch, glides, loudness levels, repetitive phonations, and forced inhalation), adduction and abduction of vocal folds, and phonatory onset and offset. Increased dynamic range allows for improved viewing quality and increased accuracy of automated image analyses. Due to insufficient clinical experiments with HSV, the necessary requirements for these factors are not yet known.

Of all factors, the color, temporal resolution and dynamic range are limited by the sensitivity of the camera sensor. The spatial resolution, temporal resolution, dynamic range and color are in a reciprocal relationship since they share the same hardware bandwidth and memory resources. The sample duration depends on the memory available, while the view angle depends on the spatial resolution and the optics installed. Therefore, the improvement of the HSV technology depends on three factors: sensitivity, hardware speed, and memory size. The most important, but most challenging factor is the sensitivity, due to the limitations of CMOS technology and the lack of demand for high sensitivity from the traditional market sectors using high-speed cameras. That is, the market share of laryngeal imaging alone is relatively too small to motivate the high-tech industry to invest in advanced research.

Methodological Challenges

The HSV recordings contain exceptionally rich information about the vibratory behavior of the vocal folds that is clinically relevant. The true spatial-temporal content allows for the accurate visualization and measurement of the mucosal wave, regularity, symmetry, vocal fold edge, glottal closure, amplitude, and open quotient.

HSV allows for the visualization of nonstationary laryngeal dynamics, such as onsets, offsets and breaks.

However, the HSV visualization tools are not effective for clinical use. It takes a very long time to view the HSV playback. For example, 10 s of HSV data recorded at speed of 10,000 fps would require 2 h and 46 min to view the whole recording at a speed of 10 fps. It is challenging to perceptually judge all clinically relevant features simultaneously from the HSV playback. Some of these features cannot be perceived without special tools, such as kymography. Several new image processing techniques [7, 8, 10–15] allow for the registration and objective measurement of vocal fold features. However, the questions of selecting the most representative segments for applying these techniques, demonstrating their validity and reliability, and making them intuitive for clinical application, remain open. Additionally, artifacts such as endoscopic motion and light reflections impede the application of many of these image processing techniques.

Lack of Clinical Relevance Data

Currently, clinical protocols using HSV are not widely available. The clinical relevance of HSV is dependent on the ability of the clinician to differentiate norm from pathology in the images obtained. At present, clinicians are not able to effectively use HSV to this end. One main limitation to the clinical use of HSV is the lack of norms. Given that this technique is relatively new and possibly visualizes new features, clinicians need to know whether what they see falls within the confines of normal vocal fold vibratory behavior. Until this is known, there is a risk that many vocally normal persons may be considered vocally disordered. Even though this equipment provides more details of vocal fold physiology, clinicians have grown accustomed to utilizing and interpreting stroboscopy and the norms currently employed for laryngeal visualization are based on stroboscopy. There are no formal comparative studies of HSV and stroboscopy, other than those from our team detailed below, available in the literature. Thus, one step in developing a clinically useful HSV system is to compile norms.

In addition, there is only anecdotal evidence of the superiority of HSV. No clinical data demonstrating the unique benefits of HSV is available from formal class I (randomized, controlled clinical trial) or class II (observational, clinical study with concurrent controls) studies. Until the practicality, validity, reliability, and clinical relevance of HSV are formally studied, voice specialists will not be willing to change their methods for evaluating the vibratory behaviors of the vocal folds.

Recent Progress and Prospects

The USC Voice and Speech Lab attempted to address all of the above challenges by coordinating the effort of a multidisciplinary and multi-institutional team. While this effort is ongoing, the following is a summary of what we have achieved and learned thus far.

Color High-Resolution HSV

In our effort to significantly improve the HSV technology, we integrated a color high-resolution HSV system. First, we identified a small company, Vision Research Inc. (Wayne, N.J., USA), which had advanced high-speed imaging hardware mainly designed for the defense, automotive and movie industries. In January 2003, we realized the first color 2,000 fps HSV recordings of vibrating vocal folds. Although at that time the Phantom cameras were sufficiently advanced to allow for frame rates of up to 50,000 fps, they were lacking the high standard of sensitivity needed for laryngeal HSV. The typical applications of high-speed digital imaging do not require high sensitivity as recording is usually made at daylight or at high-intensity artificial light. We demonstrated to the company the need for high sensitivity in our application and facilitated its realization. Our preliminary clinical findings using the color HSV available at the time were presented to the research community in 2005 [16].

More recently, Vision Research integrated a new self-resetting CMOS (SR-CMOS) sensor with custom features allowing for a very high dynamic range. The new Phantom v7.3 camera has a sensitivity of 4,800 ASA, dynamic range of 14 bit per channel (42 bit RGB), spatial resolution of 800 × 600 pixels, maximum speed of 190,476 fps (6,688 fps at full resolution), minimum exposure time of 1 μs, and maximum memory size of 16 GB. That camera, equipped with a 70° rigid endoscope (model 9106, KayPENTAX, Lincoln Park, N.J., USA) and a 300 W constant xenon light source (model 7152, KayPENTAX) allows for HSV with the following characteristics: true color 24 bit RGB, temporal resolution up to 10,000 fps, spatial resolution up to 800 × 600 pixels (640 × 480 at 10,000 fps), sample duration up to 32 s, and view angle same as in stroboscopy. The monochromatic version of the same camera allows for 14-bit gray-scale HSV recordings at temporal resolution of up to 30,000 fps.

The second challenge of this implementation was the weight of the camera. The ultra-high speeds require all hardware, including the memory, to be physically located inside the body of the camera. In order to compensate for

the weight (3.18 kg), we attached the device to a camera crane model CamCrane 200 (Glidecam Industries, Inc., Kingston, Mass., USA). The camera weight was balanced, to appear weightless to the operator, while allowing most degrees of motion freedom by using a ballhead. Other than the weightlessness, the crane was found to significantly reduce the endoscopic motion and tilt, while also introducing a comfortable system to operate. Based on that experience, we recommend using a camera crane regardless of the weight of the camera.

Endoscopic Motion Compensation for HSV

During the initial development of the HSV image processing techniques, we became aware of the necessity of reliable endoscope motion compensation (MC) [7]. Endoscopic motion affects the time alignment of the HSV image pixels, making it difficult to track the dynamic characteristics of the laryngeal structures. Thus, techniques for estimating the endoscopic motion were developed in order to factor out the motion from the analyses, or to compensate for it in the image. For instance, when distorted by motion, DKG displays intermittent changes of the voicing pattern obstructing interpretation and further analysis. The endoscopic motion should be removed for DKG to exhibit the true pattern corresponding to the selected scan line across the vocal folds [17].

We developed and evaluated a method for automatic endoscopic MC specialized for laryngeal HSV of sustained phonation. Specifically, the left-right and posterior-anterior endoscopic shifts were addressed. The method is based on the hypothesis that the difference between endoscopic and vocal fold dynamics is sufficient for accurate estimation of the endoscopic motion relative to the vocal folds. The results of this study [17] demonstrated that subpixel endoscopic MC is a valid, reliable and accurate technique. The MC method based on minimizing the L_2 distortion of the smooth time differential of HSV using convolution showed remarkable robustness. More recently, additional research [18, 19] allowed for the speed and performance optimization of the MC method. Several alternative MC algorithms were developed. The best results were obtained with new fast Fourier transform-based algorithms [18, 19], which performed up to 120 times faster than the original convolution-based algorithm.

Facilitative Playback Techniques for HSV

The improvement of the HSV camera technology is essential for the accurate registration of the biomechanical information of vocal fold movement. Advanced image processing techniques will allow for automated analyses and measurements. But the biggest advantage of HSV is the visual presentation of movement of an anatomic structure that humans, especially the skilled clinicians, understand best. Clinicians, including speech-language pathologists, phoniatricians and laryngologists, have been trained to interpret images, thus, there is no good reason for substituting the rich visual content with scalar numbers, curves and shapes that are less intuitive. Many such scalar measures and graphic presentations will likely be useful when the content is not visually accessible or when results need to be summarized for comparative purposes. However, the visual presentation of the essential data is what clinicians need most. That is, the relevant data should be presented in the most intuitive form by preserving the spatial coordinates to the maximum extent possible.

An important enhancement of our new protocol for evaluation of vocal fold behavior is the development of tools allowing enhanced visualization of the content of HSV with the possibility of improved quantification. This visualization is achieved by displaying laryngeal characteristics not easily perceivable by viewing the conventional HSV playback. We developed a set of facilitative visual playbacks [20] that are used in addition to the HSV playback to evaluate laryngeal function.

A HSV recording can be described as a function $f(x,y,t)$, which is a series of frames $f_k(x,y)$, $k = 1,2, ..., K$, in time, where K is the number of frames. For instance, capturing 10 s of HSV data at a rate of 10,000 fps would result in $K = 100{,}000$ frames. The first dimension, seen as the width of the image, is typically referred to as the 'left-to-right' axis, while the second, height, is referred to as the 'posterior-anterior' axis.

DKG Playback. The normal sequence of viewing $f(x,y,t)$ is in time t, termed *HSV playback*. Watching a sequence of frames $f_k(x,y)$, provides a good comprehension of space relations and topology that slowly change in time. DKG corresponds to viewing frames $f_y(x,t)$ of HSV along the posterior-anterior axis y, $y = 1,2, ..., N$. DKG frames can be viewed as a movie sequence that plays from posterior to anterior, replacing the time axis with the height axis. Such view is termed *DKG playback* (fig. 2). We have used facilitative visual playbacks both clinically and as part of research studies [21–25]. With minimal effort, we were able to interpret the valuable and synthesized information they provide. The posterior-to-anterior DKG playback was found useful for demonstrating the change of the dynamic characteristics while viewing damaged tissues, such as lesions, scars, and discolored

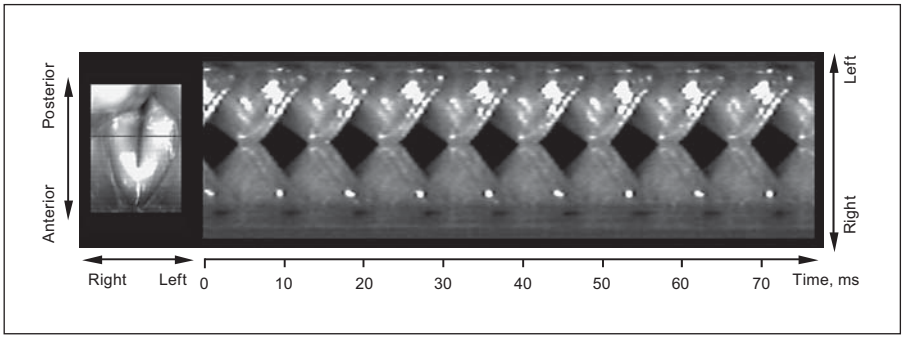

Fig. 2. A medial position frame of a DKG playback, which is a movie playing from posterior to anterior. The image on the left shows the line being scanned across the glottis. On the right, the corresponding kymographic image is shown.

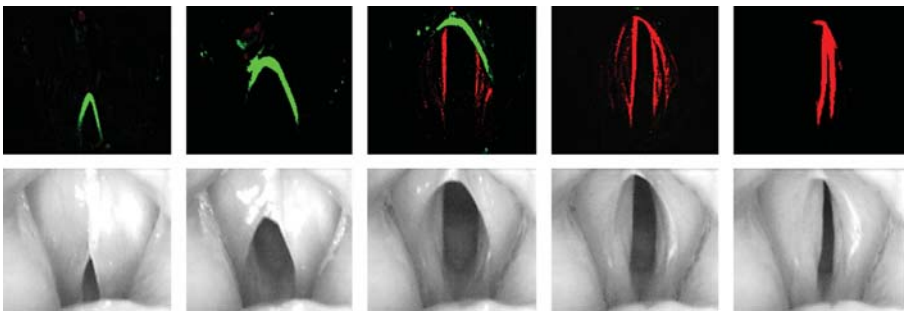

Fig. 3. Frames of mucosal wave playback (top) at different phases of the intraglottal cycle and the corresponding HSV frames (bottom). The opening phase motion of the mucosal edges is encoded in shades of green (light gray in print version), and the closing phase motion is displayed in red (dark gray in print version).

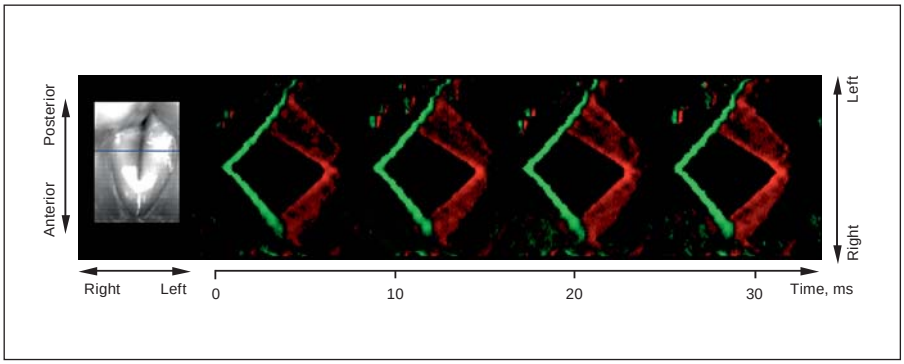

Fig. 4. A medial position frame of a MKG playback. This type of display allows for the temporal representation of the dynamics of the mucosal edges during glottal opening and closing in consecutive glottal cycles during sustained phonation. The MKG image brightness relates to the speed of motion of the mucosal edges, and the color shows the phase of motion (in print version, left half of diamond shape = opening, right half = closing). The mucosal wave extent appears as a double-edged or thicker curve during the closing phase.

areas. Dynamic changes due to stiffness of the tissue, shown as a movie, may also help to reveal the nature of lesions (i.e. cysts vs. polyps). It is essential to point out the importance of endoscopic MC in ascertaining time alignment of the anatomic structures for valid DKG representations.

Mucosal Wave (MW) and Mucosal Wave Kymography (MKG) Playbacks. Another interesting concept is modifying the HSV image sequence into a series of frames $f_k^{MW}(x,y)$, in which the pixel intensity f^{MW}, $f^{MW} = -F/2, \ldots -1,0,1, \ldots, F/2 - 1$, encodes the motion of the glottal and the mucosal edges, and the color encodes the direction of the motion. As illustrated in figure 3, in the *mucosal wave playback*, the brightness relates to the speed of motion, and the color shows the phase of motion, i.e. opening is green and closing is red. *MKG playback* corresponds to viewing kymographic frames $f_k^{MW}(x,y)$ of mucosal wave along the posterior-anterior axis $y = 1,2, \ldots,N$ (fig. 4). The mucosal wave and MKG playbacks have a substantial clinical potential as facilitative visual techniques since they allow for the assessment of the fine detail of the mucosal wave including the propagation of the mucosal edges during the opening and closing glottal phases.

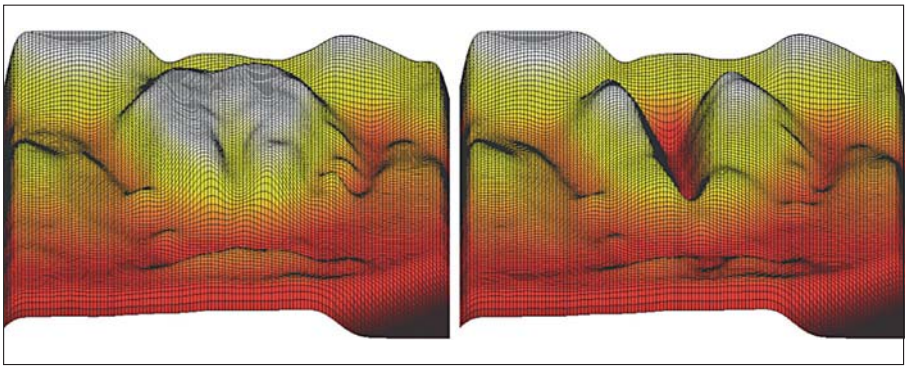

Fig. 5. 3D graphic representations of a closed and open phase within a single glottal cycle. The vertical component of the glottal movement is perceivable, particularly when sequences of 3D frames are played. The vertical movement, even though not in a linear fashion, is encoded in the HSV pixel intensity.

Three-Dimensional (3D) Playbacks. In the 3D playback, the intensity of each image pixel is represented in a separate dimension rather than in shades of gray in order to facilitate viewing of the vertical (inward-outward) motion of the larynx during phonation (fig. 5). This playback is based on the assumption that the pixel intensity is a quadratic function of the distance from the vocal folds to the tip of the endoscope, where the light source and camera lens are located. Thus, the pixel intensity is changing with the change of proximity of the endoscope, providing information regarding vertical motion of the vocal folds. While the first results [24] of using the 3D playback are encouraging, independent validation of the technique has not been done yet. In addition, the success of light reflection removal is essential for the 3D playback, since such reflections appear as large peaks obstructing the perception of the vertical laryngeal motion.

Stroboscopic Simulation Audio (SSA) Playbacks. One of the advantages of stroboscopy is the ability to play the illusionary slow motion of the vocal folds in real time along with the audio recording of the patient's voice. Such playback helps the clinician to better link the phonatory task to the images and provides valuable information regarding the normality of the vocal folds' vibratory behavior. HSV playback is typically viewed at frame rates that are hundreds of times slower than originally recorded, therefore viewing it is a time-consuming operation and sound cannot be played synchronously. However, since HSV contains all necessary information, it is feasible to produce a stroboscopy-like playback (termed *SSA playback*) allowing clinicians to 'quickly preview' the HSV while listening to the sound. SSA playback feels exactly like viewing a stroboscopic recording, but also provides a better slow motion illusion, including the ability to visualize intracycle slow motion for aperiodic vibration. This enhancement is due to the fact that the synchronization of the phase delay in SSA is based on the actual glottal cycle extracted from the HSV image rather than on the prediction of a glottal cycle from the previous cycles made from indirect signals such as acoustic or EGG waveforms.

Segmentation and Clinical Presentation of HSV

HSV contains rich temporal and spatial information about the laryngeal dynamics. However, it takes a very long time to view the HSV playback and it is impractical to use kymography and other facilitative playbacks on lengthy HSV samples. Luckily, over 90–95% of the information available through HSV is dynamically redundant. Therefore, during sustained phonation, viewing shorter segments inclusive of just a couple of glottal cycles can be representative for the relevant information in the HSV sample. Nevertheless, selecting the most representative segments for viewing is a time-consuming and tedious task, yet the criteria for such selections are subjective and nonuniform across patients and over time. Therefore, even with the technological and methodological enhancements described above, the clinical implementation of HSV is not practical, objective nor reliable.

There is a need for a reliable cycle-to-cycle similarity measure that can automatically and objectively determine the information that is redundant and the segments that are representative of that information. This is a task requiring automatic image segmentation. Segmentation can allow for navigating through the rich HSV content by mapping the specific areas of interest. Different strategies and criteria can be used when applying automatic segmentation, i.e. dispersion of the fundamental frequency (F_0) contour, the symmetry line, the glottal width, the open quotient, or all in combination.

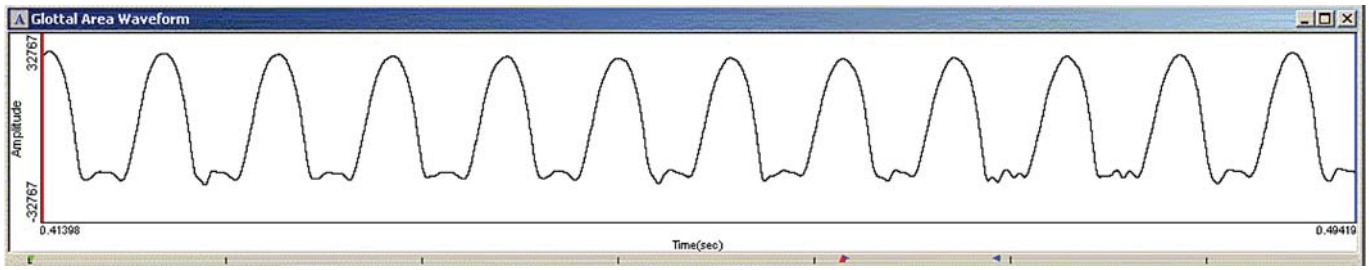

Fig. 6. GAW extracted through the second central moment of intensity of the HSV image.

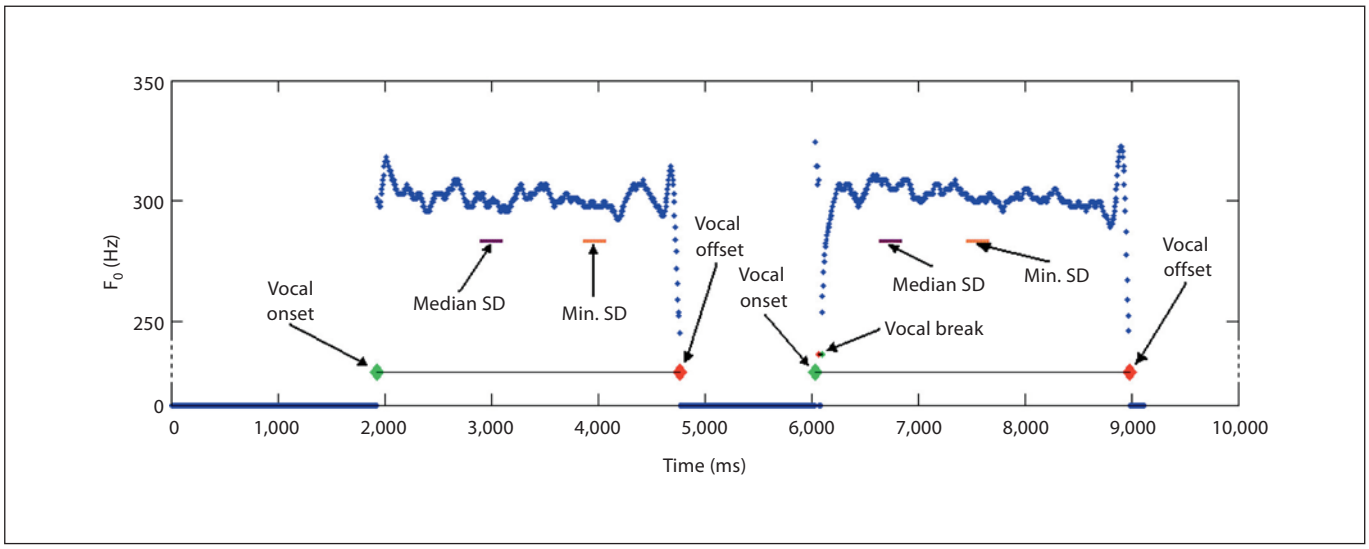

Fig. 7. HSV segmentation based on tracking the dispersion of the F_0 contour as computed from the GAW.

We have developed [20] and used successfully [21–26] a simple segmentation method based on tracking the F_0 contour of the glottal area waveform (GAW). The GAW (fig. 6) is estimated as

$$A(t) = \int I_2(y,t)dy, \quad (1)$$

where:

$$I_2(y,t) = \sqrt{\frac{\int f(x,y,t)x^2\,dx}{\int f(x,y,t)dx} - I_1^2(y,t)} \quad (2)$$

is the second central moment of intensity, or second moment of the HSV image, which is an estimate of the glottal width waveform [7], and

$$I_1(y,t) = \frac{\int f(x,y,t)x\,dx}{\int f(x,y,t)dx} \quad (3)$$

is the centroid of intensity, or average first moment of the HSV image, which provides an estimate of the glottal symmetry waveform [7].

The F_0 contour of the GAW is then computed (fig. 7) using an autocorrelation-based pitch extractor with a 40-ms Hamming window and a shift of 5 ms. The F_0 contour allows for the determination of the beginnings and ends of the periodic oscillations of the GAW, signifying the vocal onsets, offsets and breaks. Segments of 200 ms are then automatically chosen (±100 ms around each beginning point of an onset, offset or break) (fig. 7). The segments are saved as short HSV playbacks, and are consequently converted to DKG playbacks (fig. 8), allowing for the viewing of these transitional events in spatial and spatial-temporal domain, respectively.

In a following procedure, the standard deviation is computed for each 200-ms segment around each F_0 value

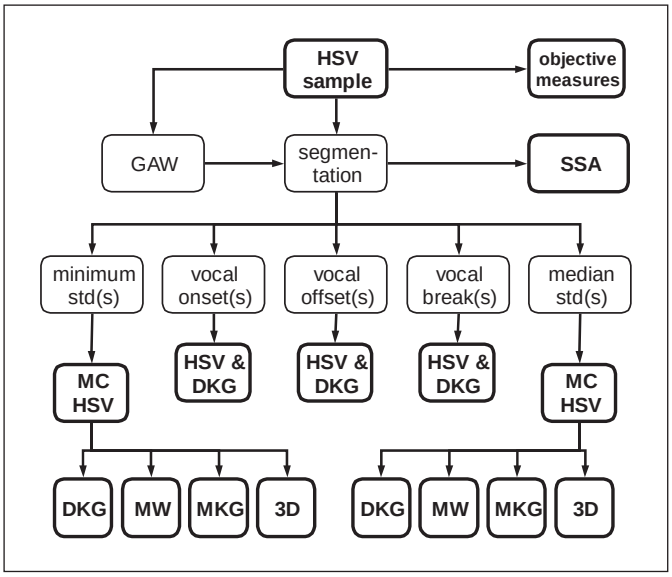

Fig. 8. Block diagram of the segmentation algorithm allowing to represent the HSV sample via facilitative playbacks (SSA, HSV, DKG, MW, MKG and 3D) and objective measures. The tokens of sustained phonation are first subjected to MC. std = Standard deviation.

(±100 ms) in the contour, resulting in a series of standard deviation values. The global minimum and the median values for each sustained phonatory portion of the standard deviation series are then computed. In this, the minimum standard deviation values signify the most stable part of phonation (in terms of F_0), while the median values are representative of the most typical parts of the vibratory pattern to be visually judged. After applying MC for the selected 200-ms segments corresponding to the minimum and median standard deviation, the HSV, DKG, mucosal wave, MKG and 3D playbacks are produced and saved as new movie sequences (fig. 8).

Finally, the SSA playback of the whole HSV sample is produced and scalar objective measurements [7] are computed (fig. 8).

This approach to segmentation is, of course, limited. It is based on tracking only the dispersion of GAW periodicity. Thus, it is assumed that the changes in vibratory dynamics always result in a change of the F_0 of the glottal area. However, not every change of vocal fold dynamics affects the periodicity. For example, a transition in glottal symmetry or open quotient may or may not influence the F_0. Therefore, some nonredundant and clinically relevant information may be skipped, as it may appear redundant. The need for more complex segmentation approaches is obvious. Although simple and limited, the presented segmentation was found reliable, practical, and effective in eliminating subjective biases when selecting the pertinent information in a large HSV sample.

Normative Data for Quantitative Assessment of Laryngeal Function via HSV

With the additional parameters available through the use of HSV, new decisions must be made in reference to normality of vocal fold movement. For example, should clinicians accept any asymmetry as abnormal and possibly pathological, or is there a degree of asymmetry that is within normal limits? As part of a recently finalized research effort, we developed a clinical protocol for HSV, implemented segmentation and facilitative playback techniques [20], collected a database of monochromatic and color HSV images from 52 vocally normal participants, and analyzed the collected data to establish preliminary normative values for HSV. In this study, we assessed the new technique by comparing its results with the standardized methods that have been well established. A summary of results from the visual-perceptual ratings of mucosal wave [21], regularity [22], open quotient [23], symmetry [23], and vertical motion [24] from comfortable and pressed phonations of normophonic speakers recorded using stroboscopy and HSV follows.

Mucosal Wave. There are three conclusions from this study [21]: atypical mucosal wave was abundant for normophonic speakers; variations in ratings were observed across visual playback techniques, and the HSV frame rate influenced visibility of mucosal wave. (1) The results of this study reinforce the presence of atypical magnitude and symmetry of the mucosal waves in the vocal fold vibration of normophonic speakers. Thus, caution should be used when determining the abnormality of mucosal wave variations during clinical visualization procedures, especially when applying the norms from stroboscopy to HSV. (2) The variation of ratings across the HSV and HSV-derived playbacks demonstrates the strength of utilizing different views providing a balance between specificity and sensitivity. This suggests that it may be beneficial to use HSV and HSV-derived playbacks in ensemble to maximize the visualization of mucosal wave. (3) An important conclusion of this investigation is the finding that the temporal resolution of 2,000 fps (typically used in commercial systems) is insufficient to record the intracycle information necessary to assess features of mucosal wave when the frequency of vocal fold vibration is high. This conclusion is based on findings of reduced mucosal wave from the habitual phonations of females with F_0

above 200 Hz. It is known that mucosal wave is reduced at high pitch. However, a follow-up comparison of high-pitched phonations recorded at 2,000 and 10,000 fps showed that the reason for reduced mucosal waves in females was the temporal resolution of HSV. Future studies should address the differences in information provided via full view and kymographic techniques, and the normality of variations in extent and symmetry of mucosal wave magnitude.

Regularity. Through the visual rating of period and glottal width regularity [22] of vocal fold vibration, using stroboscopy and HSV-based facilitative playbacks, a larger number of glottal width than period irregularities were noted. A similar difference was realized through objective measures. If the relative lack of period irregularities is further substantiated, this finding would change our understanding of indirect measures of frequency perturbation [26]. Additionally, approximately twice as many cases of glottal width irregularity were noted for pressed versus habitual phonations. This finding supports the notion of pressed phonation relying on a less typical, and possibly less stable vibratory pattern. The increased irregularities noted from the kymographic playbacks stress the importance of kymography in the voice evaluation arsenal. The need to further evaluate and compare regularity with a larger normative sample and populations of persons with voice disorders is clear. Also apparent is the continued need to validate the commonly utilized indirect clinical methods through the direct techniques of HSV and DKG. The differences between visual ratings and objective measures demonstrate the importance of quantification to evaluation of vocal fold vibratory irregularities.

Open Quotient. The majority of normophonic speakers exhibited open quotients [23] between 40 and 60% for both modal and pressed phonations when rated and measured from laryngeal videoendoscopic recordings. When variations occur, for modal phonation it is most likely that a trend towards increased open quotient is noted. Conversely, variations typically take the form of decreased open quotient for pressed phonation productions. Discrepancies between perceptual and objective measures of open quotient reveal the ability of future quantitative analysis techniques to strengthen both research and clinical applications of open quotient measures. Future investigations should compare these findings to those from a variety of pathologies and further study the relationship between direct evaluations of open quotient and information gained through indirect techniques.

Symmetry. Investigations of horizontal (left-to-right) symmetry revealed more instances of asymmetry than symmetry across all playbacks [23]. In comparison, there was an even larger prevalence of asymmetrical ratings for longitudinal (posterior-to-anterior) symmetry. An increased percent of cases were rated as symmetrical during pressed versus comfortable habitual phonations, which could reveal physiological differences between the two types of phonation.

Vertical Motion. The study of vertical motion [24] of vocal fold vibration revealed an increased magnitude of vertical motion during pressed phonations as compared to comfortable habitual phonation. This finding is in support of the whiplash theory [27] of tissue damage.

The research focus of establishing norms for HSV is being continued in the USC Voice and Speech Lab through three means. One, we are continuing the collection of normative data while also improving the technology and methodology of HSV. Two, we have begun a trial application of the preliminary norms established in the Lab to pathological cases. Three, we are studying nonstationary laryngeal dynamics, mainly throat clearing and cough, glottal attack, and vocal onset and offset, for the establishment of norms and elucidation of common clinical phenomena.

Significance of HSV in Phonomicrosurgery

The need to optimize HSV for clinical and research use is actually more pressing than ever because of the recent acceleration of efforts to develop surgical methods and bio-implants for repairing damaged vocal fold superficial lamina propria, a major factor in most voice disorders. These efforts to restore the delicate biomechanical properties of the superficial lamina propria need the type of increased accuracy that HSV can potentially provide to assist with more precisely defining/mapping specific damaged areas to target for repair and/or reconstitution of the superficial lamina propria pliability, and for accurately assessing the impact of such interventions on the fine temporal details of vocal fold vibration.

The new color high-resolution HSV system was implemented at the Center for Laryngeal Surgery and Voice Rehabilitation at the Massachusetts General Hospital. It is expected that the higher temporal resolution and registration of true intracycle glottal dynamics by HSV will provide more accurate descriptions of the underlying pathophysiology of disordered voice production than is possible with stroboscopy. Such increases in accuracy will lead to important refinements in the assessment and diagnosis of vocal pathology. Quantitative and qualita-

tive approaches are being used primarily focusing on comparisons of HSV measures with intraoperative characterization of vocal fold pathology, both in terms of preoperative assessments and with respect to relative changes in measures postoperatively (i.e., comparisons of presurgery and postsurgery assessments). If this study shows positive results, we believe that the HSV methods will be widely adopted in clinical practice to improve the assessment of voice disorders, and will become a particularly valuable and necessary tool for the development, evaluation, and continued improvement of advanced phonomicrosurgical procedures.

Conclusions

It appears that the future of HSV as a clinical technique is not in supplementing stroboscopy, but rather in supplanting it. Thus, along with demonstrating new clinically useful features, a successful clinical HSV approach should possess all of the useful features of stroboscopy. We have demonstrated that the latter is achievable with the appropriate technology and methodology. Further research is necessary to establish quantitatively the clinical requirements for HSV camera specifications, i.e. necessary temporal and spatial resolution, dynamic range, memory size, and view angle, which will become possible by identifying and studying the clinically relevant spatial-temporal features. An important area for research in the methodology area is on ascertaining effective and reliable segmentation methods. In addition, there is a need to develop new, highly intuitive facilitative playback techniques, while the existing ones should be further validated and enhanced. A uniform rationale for HSV-derived objective measurements is highly needed, possibly taking advantage of the latest knowledge in nonlinear dynamics. The norms for vocal fold vibratory features should be further studied using larger samples and more advanced technology and methodology. It is especially important to define norms for nonstationary features uniquely available through HSV, i.e. for vocal onset, offset, breaks, and throat clearing and coughing. Comprehensive comparative studies between HSV and stroboscopy are essential. Special emphasis should be placed on evidence-based studies of the HSV relevance for the enhancement of voice evaluation and therapeutic approaches. In that, the most valuable demonstration of new clinically useful features using HSV is likely to come from new clinical areas, in which stroboscopy had limited or no utilization. One such new area could be the refinement of the surgical methods and bio-implants for repairing damaged vocal fold superficial lamina propria.

Acknowledgments

This research was funded by an R&PS grant 11560-KA01 from the University of South Carolina Research Foundation and a research grant from the Institute of Laryngology and Voice Restoration. The authors thank the following researchers for their contributions in the conceptualization and/or execution phases of this research: Dr. Tomasz Zielinski and Szymon Cieciwa from the Department of Instrumentation and Measurement, AGH University of Science and Technology, Krakow, Poland; Dr. Robert Orlikoff from the Department of Speech-Language Pathology, Seton Hall University, South Orange, N.J.; Dr. Steven Zeitels from the Center for Laryngeal Surgery and Voice Rehabilitation, Massachusetts General Hospital, Boston, Mass.; Dr. Lucinda Halstead from the Evelyn Trammell Institute for Voice and Swallowing, Medical University of South Carolina, Charleston, S.C.; Drs. Darrell Klotz and N. Neil Howell from Charlotte Eye Ear Nose and Throat Associates, Charlotte, N.C., and Cara Sauder from the Voice Disorders Center, University of Utah, Salt Lake City, Utah.

References

1 Casiano RR, Zaveri V, Lundy DS: Efficacy of videostroboscopy in the diagnosis of voice disorders. Otolaryngol Head Neck Surg 1992;107:95–100.
2 Eysholdt U, Tigges M, Wittenberg T, Proschel U: Direct evaluation of high-speed recordings of vocal fold vibrations. Folia Phoniatr Logop 1996;48:163–170.
3 Rees M: Harshness and glottal attack. J Speech Hear Res 1958;1:344–349.
4 Schutte HK, Švec JG, Šram F: First results of clinical application of videokymography. Laryngoscope 1998;108:1206–1210.
5 Švec JG, Šram F, Schutte HK: Videokymography: a new high-speed method for the examination of vocal-fold vibrations. Otorhinolaryngol Foniatrie 1999;48:155–162.
6 Qiu Q, Švec JG, Schutte HK: New generation videokymography – simultaneous structural and functional imaging of the vocal folds. Proc 7th Int Conf Adv in Quant Laryngol Voice and Speech Res AQL, University of Groningen, Groningen, 2006.
7 Deliyski D, Petrushev P: Methods for objective assessment of high-speed videoendoscopy. Proc 6th Int Conf Adv in Quant Laryngol Voice Speech Res AQL, Universitätsklinikum Hamburg-Eppendorf, Hamburg, 2003, vol 6, pp 1–16.
8 Tigges M, Wittenberg T, Mergell P, Eysholdt U: Imaging of vocal fold vibration by digital multi-plane kymography. Comput Med Imaging Graph 1999;23:323–330.
9 Farnsworth DW: High-speed motion pictures of the human vocal cords. Bell Lab Rec 1940;18:203–208.

10 Schwarz R, Hoppe U, Schuster M, Wurzbacher T, Eysholdt U, Lohscheller J: Classification of unilateral vocal fold paralysis by endoscopic digital high-speed recordings and inversion of a biomechanical model. IEEE Trans Biomed Eng 2006;53:1099–1108.
11 Lohscheller J, Dollinger M, Schuster M, Schwarz R, Eysholdt U, Hoppe U: Quantitative investigation of the vibration pattern of the substitute voice generator. IEEE Trans Biomed Eng 2004;51:1394–1400.
12 Neubauer J, Mergell P, Eysholdt U, Herzel H: Spatio-temporal analysis of irregular vocal fold oscillations: biphonation due to desynchronization of spatial modes. J Acoust Soc Am 2001;110:3179–3192.
13 Döllinger M, Braunschweig T, Lohscheller J, Eysholdt U, Hoppe U: Normal voice production: computation of driving parameters from endoscopic digital high speed images. Methods Inf Med 2003;42:271–276.
14 Döllinger M, Tayama N, Berry DA: Empirical eigenfunctions and medial surface dynamics of a human vocal fold. Methods Inf Med 2005;44:384–391.
15 Yan Y, Ahmad K, Kunduk M, Bless D: Analysis of vocal-fold vibrations from high-speed laryngeal images using a Hilbert transform based methodology. J Voice 2005;19:161–175.
16 Deliyski D, Shaw H, Gerlach T, Martin-Harris B, Halstead L, Sauder C, Evans M: Clinical application of color high-speed high-resolution videoendoscopy: system design and preliminary data. 34th Symp Voice Foundation: Care of the Professional Voice, Philadelphia, June 2005.
17 Deliyski D: Endoscope motion compensation for laryngeal high-speed videoendoscopy. J Voice 2005;19:485–496.
18 Cieciwa S, Deliyski D, Zielinski T: Fast FFT-based motion compensation for laryngeal high-speed videoendoscopy. Proc 4th In Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications MAVEBA, Florence, 2005, vol 4, pp 129–132.
19 Deliyski D, Cieciwa S, Zielinski T: Fast and robust endoscopic motion estimation in high-speed laryngoscopy. Proc 7th Int Conf Adv in Quant Lar Voice and Speech Res AQL, University of Groningen, Groningen, 2006.
20 Deliyski D, Shaw H, Martin-Harris B, Gerlach T: Facilitative playback techniques for laryngeal assessment via high-speed videoendoscopy. 34th Symp Voice Found: Care of the Professional Voice, Philadelphia, June 2005.
21 Shaw H, Deliyski D: Mucosal wave: a normophonic study across visualization techniques. J Voice 2006, E-pub ahead of print.
22 Shaw H, Deliyski D: Period and glottal width irregularities of vocal fold vibration. 35th Symp Voice Found: Care of the Professional Voice, Philadelphia, June 2006.
23 Shaw H, Deliyski D: Symmetry and open quotient: normative high-speed videoendoscopy findings. 34th Symp Voice Found: Care of the Professional Voice, Philadelphia, June 2005.
24 Shaw H, Deliyski D: Vertical motion during modal and pressed phonation: magnitude and symmetry. Proc 4th Int Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications MAVEBA, Florence, 2005, vol 4, pp 133–136.
25 Orlikoff R, Watson B, Baken RJ, Deliyski D: Quantifying vocal attack using a glottographic estimate of vocal-fold approximation time. Proc 7th Int Conf Adv in Quant Laryngol Voice and Speech Res AQL, University of Groningen, Groningen, 2006.
26 Deliyski D, Shaw H, Orlikoff R: Jitter as an index of irregular variation in glottal area. 35th Symp Voice Found: Care of the Professional Voice, Philadelphia, June 2006.
27 Sonninen A, Laukkanen A-M: Hypothesis of whiplike motion as a possible traumatizing mechanism in vocal fold vibration. Folia Phoniatr Logop 2003;55:189–198.