Chapter 12

Integration of transnasal fiberoptic high-speed videoendoscopy with time-synchronized recordings of vocal function

Daryush D. Mehta, Dimitar D. Deliyski, Steven M. Zeitels, Matías Zañartu & Robert E. Hillman

Abstract

This chapter reports on the development of a system that integrates the capture of laryngeal high-speed videoendoscopy (HSV) through a transnasal fiberoptic endoscope with the synchronous acquisition of multi-sensor recordings of vocal function. Laryngeal HSV is achieved by the transnasal placement of a flexible fiberoptic endoscope with its eyepiece coupled to a monochromatic high-speed video camera and distal end passed through a specially-modified pneumotachograph mask. The setup includes the simultaneous acquisition of signals from a microphone, electroglottograph, accelerometer, and transducers for intraoral air pressure and airflow. Example data illustrate the ability of the transnasal HSV system to synchronously record measures of multiple vocal function parameters from speakers with and without voice disorders. Key features of the digital high-speed camera include an output signal that provides accurate time synchronization and enhanced light sensitivity to capture monochromatic video at rates over 4000 images per second. Transnasal HSV imaging can be combined with other measures of vocal function to significantly expand the potential for comprehensive investigations into phonatory mechanisms during more natural speech production tasks, particularly with respect to the role and impact of aerodynamic forces.

Keywords: *HSV, transnasal approach, fiberoptics, pneumotachograph mask, aerodynamics, vocal folds*

Introduction

High-speed videoendoscopy (HSV) is the method of choice to accurately visualize and quantify vocal fold vibratory function in human subjects that are challenging to capture using stroboscopic imaging methods. HSV has typically relied on the use of transoral rigid endoscopes that are passed through the mouth and are able to deliver the sufficient amount of illumination necessary for HSV at frame rates above 2000 Hz [1].

Coupling HSV data with simultaneous measures has enabled detailed investigations into relationships between vocal fold vibratory function and other key phonatory parameters including vocal fold tissue contact [2-3] and the acoustic characteristics of the voice [4-6]. The capability to more accurately determine the impact of vocal fold vibratory function on the acoustic characteristics of the voice is critical to advancing our understanding of phonatory mechanisms that would improve the clinical assessment and treatment of voice disorders. Thus far, such efforts have only shown mild to moderate correlations between measures of vocal fold vibration and acoustic measures related to voice quality.

More specifically, HSV-based measures of vocal fold irregularity and symmetry displayed statistically significant, but relatively low, correlations (i.e., < 50 % explained variance) with acoustic measures of perturbation, harmonics-to-noise ratio [4], spectral tilt [5], and cepstral peak magnitude [6]. These results have prompted investigators to suggest/speculate that relationships between acoustic measures of voice quality and HSV-based estimates of vocal fold vibratory function could be better explained by the capability of simultaneously measuring aerodynamic forces during phonation. They have reasoned that the addition of aerodynamic measures would, in essence, help provide the missing link between vocal fold kinematics and the aero-acoustic mechanisms involved during laryngeal sound production.

Granqvist et al. [7] developed a system for the simultaneous display and recording of transoral HSV using a rigid endoscope, the radiated acoustic waveform, oral airflow, and intraoral air pressure (during bilabial stop consonants) using a modified pneumotachograph mask. Due to technical limitations, however, synchronization errors between video (1900 frames per second) and data signals (16,000 samples per second) limited the time alignment of the recordings to 200-frame segments. The type of speech utterance that could be produced was significantly limited due to the transoral endoscope position. Rigid endoscopy through a face mask is a challenging task for both endoscopist and subject, as careful control of tongue and lip positioning is required to produce an adequate seal around the endoscope while performing tasks yielding subglottal pressure estimates.

The introduction of flexible fiberscopes for transnasal laryngeal imaging [8] has had a significant impact in voice research and clinical practice, making it possible to examine subjects with a heightened gag reflex and to observe more natural laryngeal function during a wider range of speech and non-speech tasks, as compared to transoral endoscopy that limits the speaker's vowel configuration and involves procedures that influence how the larynx functions (such as holding the extended tongue outside of the mouth). Fiberoptic technology has been coupled with simultaneous aerodynamic measurements in imaging studies utilizing stroboscopy [9-10] and standard video rate imaging [11]. In 1996 the first laryngeal HSV recordings using a transnasal flexible fiberscope were reported in vocally normal speakers [12], taking advantage of a light-intensified digital imaging system that captured vocal fold vibration at 3000 frames per second with simultaneous acoustic and electroglottographic recordings [13]. As image sensors continue to improve, capture rates achievable in a flexible HSV setup can reach 20,000 frames per second [14].

The purpose of this chapter is to describe a system that synchronously acquires HSV recordings of the vocal folds through a flexible fiberoptic endoscope along with other non-invasively derived measurements of voice production. We have used this setup to assess fluid–structure–acoustic interactions [15] and subglottal inverse filtering [16] in normal subjects and in patients with voice disorders.

Method

Flexible fiberoptic high-speed videoendoscopy

The configuration of the flexible fiberoptic HSV system is shown in Figure 1.

HSV data are acquired using the Phantom v7.1 high-speed video camera (Vision Research, Inc., Wayne, NJ), enabling monochromatic image capture at fast rates due to a sensitive CMOS image sensor. Video rates are set to 4000 images per second and can be increased with newer camera technologies. In our setup, a 45 mm–focal length lens adapter (KayPENTAX, Montvale, NJ) is used to couple the endoscope to the camera's image sensor. Spatial calibration of images in physical units can be accomplished by identifying a reference laryngeal landmark whose dimensions can be estimated independently using a calibrated endoscope system [17].

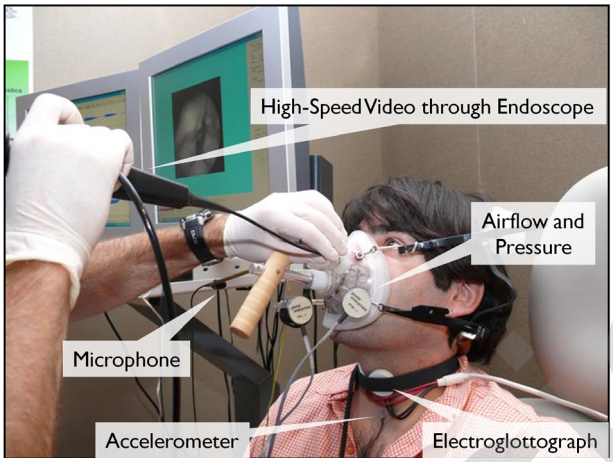

Figure 1. Illustration of transnasal fiberoptic high-speed videoendoscopy with time-synchronized measures of vocal function.

HSV data are acquired using the Phantom v7.1 high-speed video camera (Vision Research, Inc., Wayne, NJ), enabling monochromatic image capture at fast rates due to a sensitive CMOS image sensor. Video rates are set to 4000 images per second and can be increased with newer camera technologies. In our setup, a 45 mm–focal length lens adapter (KayPENTAX, Montvale, NJ) is used to couple the endoscope to the camera's image sensor. Spatial calibration of images in physical units can be accomplished by identifying a reference laryngeal landmark whose dimensions can be estimated independently using a calibrated endoscope system [17].

The flexible fiberscope has a 3.4 mm distal tip/insertion diameter and a working length of 300 mm (model FNL-10RP3; KayPENTAX). The light source contains a short-arc Xenon lamp rated at 300 watts (KayPENTAX). The fiberscope is passed through a hole in the pneumotachograph mask to access the left or right nostril. The hole is lined with an O-ring to provide a seal and slight friction so that the endoscopist can manipulate the cable through the nasal cavity [10]. Image artifacts due to light interaction with the fiber bundle must be addressed during endoscopy and/or post processing. For example, objects located farther from the center of images suffer from optical barrel distortion, and illumination through individual fibers in the bundle yield moiré patterns.

An inherent tradeoff in temporal resolution exists when utilizing a fiberoptic endoscope instead of a rigid endoscope. The use of a fiberoptic endoscope leads to significant decrease of the overall HSV temporal resolution. This is due to two factors: the reduction in light available to illuminate the vocal folds through the fiberoptic bundle versus through a glass rod; and the reduction in reflected light delivered back to the camera through the fiberoptic bundle versus through reflecting mirrors. Decreased light increases the exposure time necessary for a satisfactory dynamic range (i.e., image quality), reducing the maximum frame rate of recording. Given the same camera sensor, monochromatic cameras provide roughly a four times larger dynamic range because they do not require color filters that further reduce light to the image sensor.

The spatial resolution of an HSV system may be enhanced by a larger-diameter lens but again at the expense of a reduction in light due to the physics of lens optics. Spatial resolution is often quoted in the literature as the number of horizontal and vertical pixels of a recorded image. However, a more appropriate resolution specification is the number of pixels that actually quantize the motion of the vocal folds in the mediolateral

and anteroposterior dimensions. Decisions regarding temporal and spatial resolution depend on the goals of the study. Temporal HSV measures—such as open quotient, speed quotient, closing quotient, and phase asymmetry—require a high frame rate. Amplitude-based measures—such as glottal area, glottal width, axis shift, vocal fold displacement, and amplitude asymmetry—require good spatial resolution.

Synchronous multi-sensor data acquisition

In addition to HSV, the system employs five sensors to quantify various aspects of vocal function: acoustic microphone, electroglottograph, neck-placed accelerometer, oral airflow transducer, and intraoral pressure transducer. The acoustic signal is recorded using a condenser microphone mounted on a rod affixed to a modified Rothenberg mask. Electroglottography (Model EG2-PC; Glottal Enterprises, Syracuse, NY) provides measures of vocal fold tissue contact and includes a highpass filter to reduce noise and gross conductance changes due to neck motion. Low-frequency amplitude distortion of the electroglottographic signal [18] is compensated for by convolving the waveform with the time-reversed impulse response of the estimated hardware highpass filter (fourth-order Butterworth filter with a 3-dB cutoff frequency of 18.5 Hz).

Neck surface acceleration can provide indirect acoustic information from the subglottal system when placed appropriately [16, 19] and has shown potential for long-term monitoring of vocal function [20-22]. Thus a miniature accelerometer (BU-27135; Knowles, Itasca, IL) is included to better understand its relationship with vocal fold tissue motion and acoustics. The accelerometer is enclosed in a silicone epoxy and mounted onto a subject's neck about 1 cm above the sternal notch using Double Stick Discs (Model 2181, 3M, Maplewood, MN).

A major advantage of our system is the ability to simultaneously acquire estimates of oral airflow and intraoral air pressure using a Rothenberg mask (Glottal Enterprises, Syracuse, NY) that has been modified to allow the transnasal insertion of the flexible fiberoptic endoscope. With the scope inserted, the mask system functions as usual with one transducer providing a high-bandwidth estimate of oral airflow by measuring the pressure drop across the wire screen of the mask, while a second transducer provides an estimate of intraoral air pressure via a translabial catheter. Subglottal pressure estimates are obtained during the production of /p/ consonants as the pressure between the oral cavity and lung volume equilibrates [23].

Figure 2 shows a wiring diagram of the system that highlights the time synchronization of the video and data signals that is critical for comparing characteristics among HSV and other voice production measures from the same phonatory segment and within individual glottal cycles.

The camera's sampling rate (clock sync input) and signal sampling rate are derived from the master clock signal of the National Instruments board. The hardware clock division (e.g., dividing an 80 kHz data rate into a 4 kHz video rate) and data acquisition settings are controlled by MiDAS DA software (Xcitex Corporation, Cambridge, MA). For example, each video frame corresponds to 20 data samples at a frame rate of 4000 Hz. With both video and data rates derived from a common clock source, each HSV frame is aligned with its associated data samples by simultaneously recording a transistor-transistor logic signal from the camera that exhibits a falling edge at the end of the exposure duration of the last recorded image in the buffer.

Previous studies used a simple trigger signal to align video with simultaneously acquired data [7, 24]. The data fell out of synchrony, however, for recorded segments far from the trigger event due to small drifts in the different clocks of the video and data acquisition systems and uncertainties between the trigger signal and video frames.

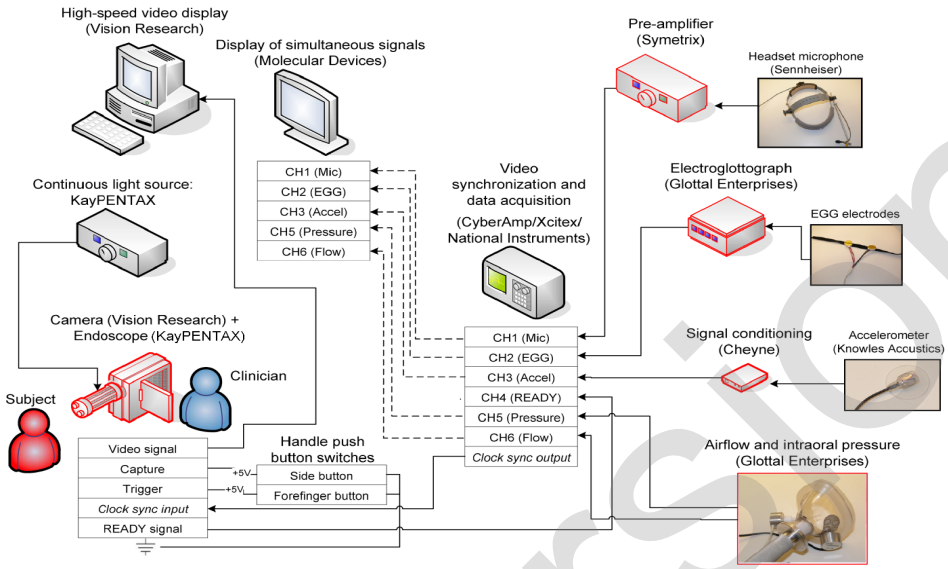

Figure 2. Wiring diagram of the HSV system with time-synchronized measures of vocal function obtained from a microphone (Mic), electroglottograph (EGG), accelerometer (Accel), pneumotachograph (for oral airflow and intraoral air pressure).

All signals exhibit physiological delays with respect to the glottis due to sensor locations. A simple method to correct for these delays to time-align all signals at the glottis is described here (individual speaker variations require subject-specific processing). The radiated acoustic signal must be shifted back in time to compensate for acoustic propagation time from glottis to microphone (~600 µs, obtained with a distance of 21 cm and a speed of sound of 34,000 cm/s). Whereas electroglottography requires no compensation, the neck-skin acceleration signal must be shifted approximately 125 µs back in time due to the position of the accelerometer approximately 5 cm below the glottis [15-16]. The oral airflow at the lips is shifted 500 µs back in time (17 cm vocal tract length and speed of sound of 34,000 cm/s). Precise alignment of the intraoral pressure signal is usually less critical because measures of interest are often static amplitudes during the closed phase of lip occlusions.

Results

Figure 3 illustrates the synchronized output of our system for an adult female sustaining the vowel /a/.

Shown are the acoustic signal, electroglottograph, neck-surface accelerometer waveform, and aerodynamic signals (oral airflow and intraoral pressure), along with selected endoscopic vocal fold images from the corresponding HSV sequence. Digital kymograms from anterior, middle, and posterior locations of the glottis are shown to illustrate the temporal resolution of the system. For this session, the video and data rates were set to 4016 Hz and 120,480 Hz, respectively (i.e., 30 data samples per video frame). Video images were cropped to 279 horizontal x 214 vertical pixels.

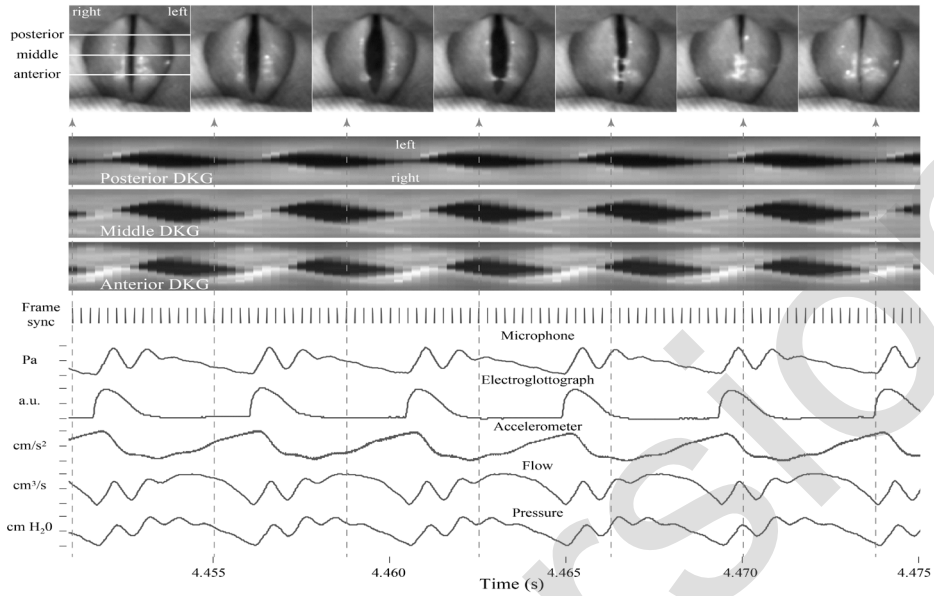

Figure 3. Illustration of high-speed videoendoscopic images, digital kymography (DKG), and time-aligned sensor signals during sustained phonation of the vowel /a/ by an adult female normal voice. Frame synchronization pulses (frame sync) indicate timing of video images (30 samples/frame). *a.u. = linear arbitrary units.*

Figure 4 shows the synchronous signal view of a recording performed on an adult female with bilateral vocal fold nodules.

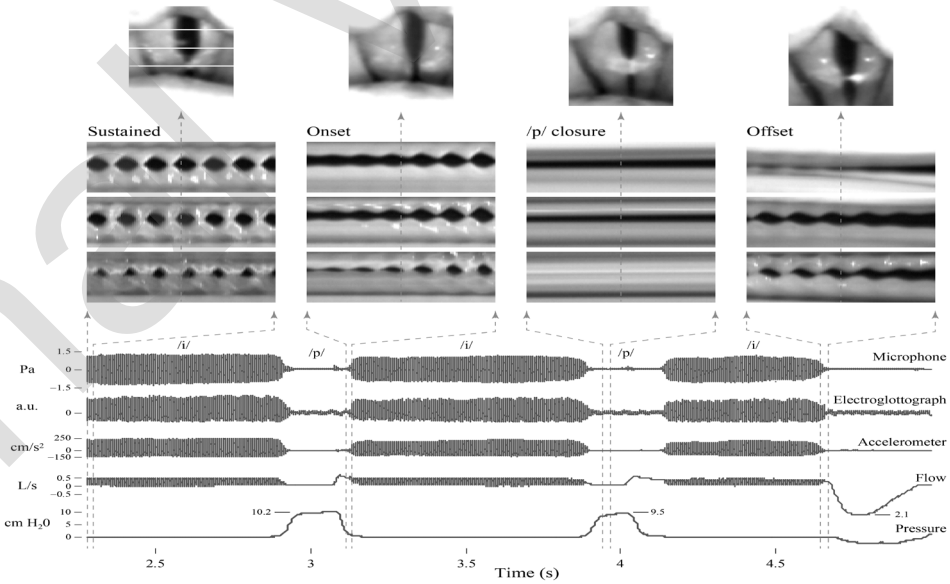

Figure 4. Illustration of high-speed videoendoscopic images, digital kymography (DKG), and time-aligned sensor signals during the production of a /pipipi/ segment by an adult female with bilateral vocal fold nodules. Video rate is 4000 Hz and sensor sample rate is 80,000 Hz. Posterior, middle, and anterior DKGs are displayed for segments during a sustained vowel /i/, at phonation onset following /p/ closure, during /p/ closure, and at phonation offset preceding an inhalation.

The subject was instructed to produce the /pipipi/ gesture to yield subglottal pressure estimates during the /p/ segments. 10,705 HSV frames were captured, corresponding to precisely 2.67625 s of signal data. Video images were cropped to 74 horizontal x 97 vertical pixels. Digital kymograms illustrate vocal fold vibratory characteristics during the following selected segments: sustained vowel, phonation onset, /p/ closure, and phonation offset. Note that the intraoral air pressure plateaus during the /p/ occlusion at approximately 10 cm H2O.

Discussion and conclusions

HSV systems have emerged as important instruments in voice research and clinical voice assessment [1, 25]. This research note reported on the development of a transnasal fiberoptic HSV system that simultaneously records time-synchronized signals from multiple sensors of vocal function. Transnasal endoscopy has several advantages over transoral endoscopy, which limits speakers to produce sustained phonation (of only one vowel) and involves procedures that influence how the larynx functions (e.g., unnaturally pulling the protruding tongue and raising the larynx). Furthermore, transnasal endoscopy also provides a means to perform laryngeal examinations on individuals who cannot tolerate transoral endoscopy due to the gag reflex.

As with any multi-sensor data acquisition system, critical specifications must be met, including HSV frame rate, synchronization accuracy, recording duration, and image resolution and quality. The system reported on in this chapter captured adequate video at 4000 frames per second. Higher sampling rates, however, can be achieved with newer camera technologies entering the market each year and are necessary to track vocal fold tissue motion at high fundamental frequencies (over 300 Hz) [14]. Care must be taken when analyzing the acoustic and airflow waveforms, which exhibit known bandwidth limitations due to the use of a flow mask [26-27]. Also, because of the possibility of leaks introduced by the insertion of the endoscope through the velopharyngeal port, additional vigilance must be exercised in monitoring intraoral air pressure signals to ensure that equilibration between oral and subglottal regions is achieved (i.e., the appearance of flat tops on the pressure signal during lip closure).

It should be noted that the more natural state of the vocal tract during transnasal fiberoptic endoscopy can actually make it difficult to gain full exposure of the glottis during phonation due to the compression of supraglottal structures, which can occur often in particular voice disorders (e.g., vocal hyperfunction). As is the case in transoral endoscopy, elicitation of the vowel /i/ can aid in raising the larynx to create better glottal exposure during transnasal endoscopy.

To date, relatively few HSV-based studies have directly investigated relationships between physiological and acoustic measures of sound production in human *in vivo* experiments. Perhaps the capability to time-synchronize recordings of video and multi-sensor data and the addition of aerodynamic measures can help spur research into this area. In addition to providing new insight into phonatory mechanisms, such work is particularly important to help guide the development of new laryngeal surgical approaches, particularly those that are being designed to alter the biomechanical properties of damaged vocal fold tissue (e.g., bio-implants) [28]. Advancements in this area clearly require a much better understanding of the impact of vocal fold vibratory function on laryngeal sound production. Finally, current videoendoscopic recording technology limits image processing to two spatial dimensions. Efforts to explain relationships between vocal fold vibration and laryngeal sound production might also benefit from the capability to capture the three-dimensional motion of the vocal folds.

A note on HSV terminology

Our group advocates for using terms such as *laryngeal high-speed videoendoscopy* to describe the application of high-speed imaging techniques to the visualization of vocal fold vibration, striking a balance between conciseness and specificity. Inclusion of the term *video*—a sequence of images that reflect the natural and observable temporal progression of a moving object—makes an important distinction in disambiguating the term *high-speed*, which by itself could refer to high-speed photography or to high-speed video. Adding *endoscopy* distinguishes the technique from other imaging modalities such as optical coherence tomography, ultrasound, and magnetic resonance imaging. Finally, the acronym HSV remains flexible so it can be qualified by other anatomical terms to describe the body part being imaged (e.g., laryngeal HSV, tongue HSV). Accurate and consistent use of terminology helps disambiguate nomenclature, provide a common reference for studies, and prevent the potential for future misconceptions.

We acknowledge we would not even be having this conversation were it not for the tireless work of voice scientists and researchers before us. It has been our pleasure to celebrate the lives of and stand on the shoulders of two giants in our field—Dr. Paul Moore and Dr. Hans von Leden.

Acknowledgments

This work was supported in part by the Eugene B. Casey Foundation, the Voice Health Institute, and NIH NIDCD grant R01 DC007640. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

References

1. Deliyski, D., et al., 2008. Clinical implementation of laryngeal high-speed videoendoscopy: Challenges and evolution. Folia Phoniatr. Logop. 60, 33-44.
2. Mecke, A., Sundberg, J., Granqvist, S., Echternach, M., 2012. Comparing closed quotient in children singers' voices as measured by high-speed-imaging, electroglottography, and inverse filtering. J. Acoustic Soc. Amer. 131, 435-441.
3. Orlikoff, R., Golla, M., Deliyski, D., 2012. Analysis of longitudinal phase differences in vocal-fold vibration using synchronous high-speed videoendoscopy and electroglottography. J. Voice 26, 816.e813-816.e820.
4. Mehta, D., Deliyski, D., Zeitels, S., Quatieri, T., Hillman, R., 2010. Voice production mechanisms following phonosurgical treatment of early glottic cancer. Ann. Otol. Rhinol. Laryngol. 119, 1-9.
5. Mehta, D., Zañartu, M., Quatieri, T., Deliyski, D., Hillman, R., 2011. Investigating acoustic correlates of human vocal fold vibratory phase asymmetry through modeling and laryngeal high-speed videoendoscopy. J. Acoustic Soc. Amer. 130, 3999-4009.
6. Mehta, D., et al., 2012. High-Speed Videoendoscopic Analysis of relationships between cepstral-based acoustic measures and voice production mechanisms in patients undergoing phonomicrosurgery. Ann. Otol. Rhinol. Laryngol. 121, 341-347.
7. Granqvist, S., Hertegård, S., Larsson, H., Sundberg, J., 2003. Simultaneous analysis of vocal fold vibration and transglottal airflow: Exploring a new experimental setup. J. Voice 17, 319-330.
8. Sawashima, M., Hirose, H., 1968. New laryngoscopic technique by use of fiber optics. J. Acoustic Soc. Amer. 43, 168-169.

9. Hertegård, S., Gauffin, J., 1995. Glottal area and vibratory patterns studied with simultaneous stroboscopy, flow glottography, and electroglottography. J. Speech Hear. Res. 38, 85-100.
10. Kobler, J., Hillman, R., Zeitels, S., Kuo, J., 1998. Assessment of vocal function using simultaneous aerodynamic and calibrated videostroboscopic measures. Ann. Otol. Rhinol. Laryngol. 107, 477-485.
11. Sundberg, J., Scherer, R., Hess, M., Müller, F., 2010. Whispering: A single-subject study of glottal configuration and aerodynamics. J. Voice 24, 574-584.
12. Maurer, D., Hess, M., Gross, M., 1996. High-speed imaging of vocal fold vibrations and larynx movements within vocalization of different vowels. Ann. Otol. Rhinol. Laryngol. 105, 975-981.
13. Hess, M., Gross, M., 1993. High-speed, light-intensified digital imaging of vocal fold vibrations in high optical resolution via indirect microlaryngoscopy. Ann. Otol. Rhinol. Laryngol. 102, 502-507.
14. Echternach, M., Dollinger, M., Sundberg, J., Traser, L., Richter, B., 2013. Vocal fold vibrations at high soprano fundamental frequencies. J. Acoustic Soc. Amer. 133, EL82-EL87.
15. Zañartu, M., Mehta, D., Ho, J., Wodicka, G., Hillman, R., 2011. Observation and analysis of in vivo vocal fold tissue instabilities produced by nonlinear source-filter coupling: A case study. J. Acoustic Soc. Amer. 129, 326-339.
16. Zañartu, M., Ho, J., Mehta, D., Hillman, R., Wodicka, G., 2013. Subglottal impedance-based inverse filtering of voiced sounds using neck surface acceleration. IEEE Trans. Audio Speech Lang Processing 21, 1929-1939.
17. Kobler, J., et al., 2006. Comparison of a flexible laryngoscope with calibrated sizing function to intraoperative measurements. Ann. Otol. Rhinol. Laryngol. 115, 733-740.
18. Rothenberg, M., 2002. Correcting low-frequency phase distortion in electroglottograph waveforms. J. Voice 16, 32-36.
19. Cheyne, H., 2006. Estimating glottal voicing source characteristics by measuring and modeling the acceleration of the skin on the neck. Proc. 3rd IEEE-EMBS International Summer School & Symposium on Med. Devices & Biosensors, 118-121.
20. Cheyne, H., Hanson, H., Genereux, R., Stevens, K., Hillman, R., 2003. Development and testing of a portable vocal accumulator. J. Speech Lang. Hear. Res. 46, 1457-1467.
21. Mehta, D., Zañartu, M., Feng, S., Cheyne II, H., Hillman, R., 2012. Mobile voice health monitoring using a wearable accelerometer sensor and a smartphone platform. IEEE Trans. Biomed. Eng. 59, 3090-3096.
22. Popolo, P., Švec, J., Titze, I., 2005. Adaptation of a Pocket PC for use as a wearable voice dosimeter. J. Speech Lang. Hear. Res. 48, 780-791.
23. Rothenberg, M., 1973. A new inverse filtering technique for deriving glottal air flow waveform during voicing. J. Acoustic. Soc. Amer. 53, 1632-1645.
24. Larsson, H., Hertegård, S., Lindestad, P., Hammarberg, B, 2000. Vocal fold vibrations: High-speed imaging, kymography, and acoustic analysis: A preliminary report. Laryngoscope 110, 2117-2122.
25. Mehta, D., Hillman, R., 2012. Current role of stroboscopy in laryngeal imaging. Curr. Opin. Otolaryngol. Head Neck Surg. 20, 429-436.
26. Badin, P., Hertegård, S., Karlsson, I., 1990. Notes on the Rothenberg mask. STL-QPSR, KTH 1, 1-7.
27. Hertegård, S., Gauffin, J., 1992. Acoustic properties of the Rothenberg mask. Speech Trans. Lab. Quart. Progress Status Report 2, 9-18.

28. Karajanagi, S., et al., S.M., 2011. Assessment of canine vocal fold function after injection of a new biomaterial designed to treat phonatory mucosal scarring. Ann. Otol. Rhinol. Laryngol. 120, 175-184.