Published in final edited form as:

Curr Opin Otolaryngol Head Neck Surg. 2008 June ; 16(3): 211–215. doi:10.1097/MOO.

Voice assessment: Updates on perceptual, acoustic, aerodynamic, and endoscopic imaging methods

Daryush D. Mehta, SM and

Center for Laryngeal Surgery and Voice Rehabilitation, Massachusetts General Hospital, Harvard-MIT Division of Health Sciences and Technology, One Bowdoin Square, 11th Floor, Boston, MA 02114, 617-643-8417

Robert E. Hillman, PhD

Center for Laryngeal Surgery and Voice Rehabilitation, Massachusetts General Hospital, Harvard Medical School/Harvard-MIT Division of Health Sciences and Technology, One Bowdoin Square, 11th Floor, Boston, MA 02114, 617-726-0220

Daryush D. Mehta: dmehta@mit.edu; Robert E. Hillman: hillman.robert@mgh.harvard.edu

Abstract

Purpose of review—This paper describes recent advances in perceptual, acoustic, aerodynamic, and endoscopic imaging methods for assessing voice production.

Perceptual assessment: Speech-language pathologists are being encouraged to use the new CAPE-V inventory for auditory perceptual assessment of voice quality, and recent studies have provided new insights into listener reliability issues that have plagued subjective perceptual judgments of voice quality.

Acoustic assessment: Progress is being made on the development of algorithms that are more robust for analyzing disordered voices, including the capability to extract voice quality-related measures from running speech segments.

Aerodynamic assessment: New devices for measuring phonation threshold air pressures and air flows have the potential to serve as sensitive indices of glottal phonatory conditions, and recent developments in aeroacoustic theory may provide new insights into laryngeal sound production mechanisms.

Endoscopic imaging: The increased light sensitivity of new ultra high-speed color digital video processors is enabling high-quality endoscopic imaging of vocal fold tissue motion at unprecedented image capture rates, which promises to provide new insights into mechanisms of normal and disordered voice production.

Summary—Some of the recent research advances in voice quality assessment could be more readily adopted into clinical practice, while others will require further development.

Keywords

voice quality assessment; perception of voice; acoustic voice analysis; high-speed endoscopic imaging; aerodynamics of voice production

Introduction

“Voice” is the sound that the listener perceives when the adducted vocal folds are driven into vibration by the pulmonary air stream. The four most common approaches for clinically assessing the various aspects of voice production include: 1) auditory *perceptual assessment* of voice quality, 2) *acoustic assessment* of voiced sound production, 3) *aerodynamic*

assessment of subglottal air pressures and glottal air flow rates during voicing, and 4) *endoscopic imaging* of vocal fold tissue vibration. This paper provides a review of recent advances in each of these four assessment modalities based on literature that has been published in the last two years.

Perceptual assessment

CAPE-V

Speech-language pathologists are increasingly being encouraged to use the new Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) for clinical assessment of voice quality. The CAPE-V was the product of an international conference sponsored by Special Interest Division 3 (Voice and Voice Disorders) of the American Speech-Language-Hearing Association in June 2002. It provides a standardized framework and procedures for perceptual evaluation of abnormal voice quality that includes prescribed speech materials and visual analog scaling of a closed set of perceptual vocal attributes: overall severity of dysphonia, roughness, breathiness, straining, pitch and loudness [1]. While the CAPE-V represents an important step toward more uniform procedures for the clinical assessment of voice quality, it was not designed to resolve all of the persistent reliability problems that become even more apparent when the exact agreement of subjective listener ratings is accurately examined (e.g., use of statistical procedures that examine the exact agreement between judgments of the same voice sample) [2].

Exploring sources of listener disagreement using synthesized speech

There are ongoing efforts to better understand and account for the unreliability of auditory perceptual voice quality ratings. In one approach aimed at modeling sources of listener variability, Kreiman, Gerratt, and Ito [3] recently used copy-synthesized vowel samples of disordered voices in an attempt to identify and quantify sources of listener disagreement. The authors found that 84% of the variance in the extent to which listeners do or do not agree could be accounted for by four factors that can be controlled for by the experimental design: (1) instability of internal memory standards for levels of a perceptual dimension, 2) ability to isolate single dimensions in a complex context, 3) scale resolution, and 4) absolute magnitude of the attribute being measured. While this work is providing extremely valuable insights into sources of listener disagreement, direct clinical application is currently limited by the substantial technical and practical challenges associated with acoustically analyzing (see *Acoustic assessment* section) and synthesizing the disordered voices being assessed.

Application of psychometric principles to improve listener agreement

In an alternative approach, Shrivastav, Sapienza, and Nandur [4] have recently shown that listener variability can be minimized by applying psychometric principles when designing the listening task. Such principles include averaging the ratings from multiple presentations of the same stimulus, with the investigators showing that a minimum of five repetitions provide the best results for judging voice quality in sustained vowels. In addition, to account for scale resolution and edge effects related to the absolute magnitude of a perceptual attribute, a standardized value for each rating is computed. The averaging and standardization procedures attempt to minimize variability and response biases of individual listeners. Although this approach allows for the experimenter to obtain more reliable perceptual ratings of natural (versus synthesized) vowel stimuli, the approach has limited clinical application due to the impracticality of taking the time to have several clinicians independently rate multiple repetitions of the same voice samples.

Acoustics assessment

The validity and reliability of acoustic measures currently used in the clinic to objectively assess voice quality (e.g., jitter, shimmer, and noise-to-harmonics ratio) are inherently limited by a reliance on the accurate determination of fundamental frequency (F0), and these measures have been further restricted to the analysis of sustained vowels. F0 can be difficult or impossible to extract in disordered voices, and sustained vowels may not be representative of vocal function or voice quality during continuous speech [5]. In response to these limitations, there have been recent attempts to develop acoustic measures of voice quality that do not rely on accurate F0 estimation and can be extended to the analysis of continuous speech. Two basic directions are prominent in recent attempts to develop acoustic measures that overcome the limitations of those currently being used in clinical voice assessment.

Cepstral-based methods

One set of approaches is based on cepstral analysis (a spectrum-type method), which is inherently attractive since the cepstrum can be computed for any segment of speech and not just for steady vowel-like sounds. Recent work has demonstrated that the cepstral peak prominence (CPP) correlates well with perceptual judgments of overall severity of dysphonia [6], and research continues to understand and improve upon technical challenges of cepstral analysis methods [7, 8].

On the less positive side, the interpretation of cepstral-based measures relative to the underlying physiology of vocal fold vibration is not as intuitive as more traditional perturbation (e.g., jitter and shimmer) and noise (e.g., noise-to-harmonics ratio) measures, which points to the need for studies that can better delineate relationships between cepstral measures and vocal fold function. There is also a need for more robust studies of how well cepstral-based measures correlate with perceived voice quality attributes to better assess the clinical potential of such measures (see *Perceptual assessment* section).

Nonlinear dynamics-based methods

The second set of recently described acoustic voice measures is based on nonlinear dynamics or chaos analysis [9], which are much more robust with respect to analyzing atypical signals (e.g., aperiodic signals from pathological voices) than the measures currently being used for clinical voice assessment. In a proof-of-concept study, Zhang and Jiang [9] recently demonstrated that nonlinear measures could better distinguish between normal and pathological voices than could the commonly-used measures of jitter and shimmer. Much more work needs to be done to understand how nonlinear measures of the acoustic signal relate to the underlying physiology of voice production and whether it will be possible to further develop such measures to differentially (and meaningfully) delineate varying levels of dysphonia severity.

Aerodynamic assessment

Phonation threshold air pressure

Since the early 1980s, clinical assessment of aerodynamic voice parameters has typically involved extracting estimates of average subglottal air pressures and glottal air flow rates from non-invasive measures of intraoral air pressures and oral air flow rates during the controlled (constant pitch and loudness) repetition of simple syllable strings. It was subsequently shown that important additional information about glottal phonatory status (including the presence of pathology) could be obtained from estimates of the minimum air pressures required to initiate the softest possible voice production—the phonation threshold

pressure. Further work using mathematical and physical laryngeal models has further demonstrated that phonation threshold pressure is sensitive to vocal fold thinning, viscous shear properties of the tissue, and vocal tract inertance [10].

Mechanical devices for measuring phonation threshold air pressure and air flow

In the standard method, subglottal air pressure estimates are based on intraoral air pressure measurements that are obtained during lip closure of the p-sound. This method works because air pressure equilibrates throughout the airway (subglottal pressure = intraoral air pressure) when the vocal folds are abducted for p-sounds produced in strings of p + vowel syllables (e.g., pa-pa-pa-pa-pa). Jiang and his colleagues have raised concerns that the accuracy of subglottal air pressure estimates may be influenced by undesirable adjustments to respiratory forces and vocal tract shapes that untrained subjects can display during the procedure. For several years, Jiang's group has reported on the development of approaches to overcome these behavioral influences using mechanical devices that interrupt the oral air flow at unpredictable times during sustained vowel production to permit air pressure estimates. More recently, technological modifications to the airflow interruption technique have resulted in systems that adopt partial flow interruption to minimize effects related to abrupt cessation [11] and an airflow redirection tube to minimize the impact of laryngeal reflexes [12]. Jiang and Tao have also recently demonstrated, in a mathematical model, that the air flow rate at phonation threshold varies systematically with changes in simulated vocal fold tissue properties, vocal tract loading, and glottal area configurations [13]. These initial results provide evidence that such a simple air flow-based measure may have some clinical utility, particularly given the relative ease of measuring oral air flow versus air pressure.

Application of aeroacoustics to voice production

There has been an apparent resurgence of interest in applying more sophisticated aeroacoustic theories and approaches to the study of laryngeal sound production. This work indicates that other aerodynamic phenomenon (e.g., vortical flow and vortex shedding), beyond what is portrayed in classic source-filter descriptions of voice production, could be contributing significantly to the sound that is produced during phonation [14, 15]. This information may ultimately have clinical significance because these higher-order aeroacoustic phenomena could play an even greater role in mechanisms of disordered voice production than in normal phonatory functions.

Endoscopic imaging

Ultra high-speed digital color imaging

There is an ongoing interest in exploring the use of high-speed imaging to supplement or replace stroboscopy in the endoscopic assessment of vocal fold vibration. This is because stroboscopy only provides a highly averaged view of the vibration pattern and is not capable of resolving detailed tissue motion within individual vibratory cycles. Digital video camera systems have recently become available with adequate light sensitivity and recording speeds to capture 4,000 to 10,000 high-resolution color images per second through a transoral endoscope [16], a substantial improvement over previous high-speed systems that produced lower quality grayscale images at slower capture rates. The new higher-speed systems provide adequate imaging for examining higher-pitched phonation (i.e., sampling a sufficient number of images per vibratory cycle) and facilitate direct correlations with recordings from other voice measurement devices that capture signals at comparable sampling rates. For example, the capability to accurately synchronize and compare vocal fold vibration captured at 10,000 images per second with the simultaneously-recorded acoustic (microphone) signal that is also sampled at 10,000 Hz promises to provide new

insights into relationships between vocal fold tissue motion and sound production (e.g., relationships between asymmetries in tissue motion and perturbations in the acoustic signal). High-speed digital imaging can also facilitate and further enhance videokymographic assessment of vocal fold vibration, allowing detailed imaging of a glottal axis perpendicular to the midline to better examine the symmetry of vibrations between the left and right vocal folds at a chosen location. Judgments of asymmetry and disordered classification schemes based on kymographic images have been recently advocated by Švec and his colleagues [18].

The recent surge of interest in developing high-speed endoscopic imaging of the vocal folds has spawned two major approaches for analyzing the massive amount of imaging data that is generated by these systems. One approach focuses more on facilitating visual inspection of important parameters in the native images (mucosal wave, symmetry of vibration, etc.). The other approach utilizes higher-order quantitative methods (e.g., Nyquist plots) to characterize features, such as glottal area variation, that are extracted from the images but not directly related back to image visualizations.

Methods to facilitate visualization of high-speed images

Deliyski and his colleagues have focused their efforts on developing software to facilitate the efficient identification and visual examination of important phonatory events (voice onset, phonatory instabilities, etc.) with additional capabilities to quantify selected vocal parameters and features (open quotient, symmetry of vocal fold vibration, etc.) [16]. Examples of two such displays that facilitate visualization of the mucosal wave—mucosal wave playback and mucosal wave kymography—are shown in Figure 1. These new image analysis tools have recently been studied, and investigations have begun to reveal that even some degree of asymmetry during vocal fold vibration can be tolerated in normal-sounding voices [17].

Quantitative approaches to assessing high-speed vocal fold sequences

Yan and colleagues have used Nyquist plots, nonlinear processing of both the glottal area extracted from high-speed imaging and from the acoustic signal, to discriminate normal versus pathological voices [19, 20]. The quantification of glottal variations at a high temporal resolution has also laid the ground work for more sophisticated computer models that can simulate asymmetric vocal fold tissue motion observed in excised larynx preparations [21] and live human subjects [22, 23]. Developing more accurate and robust models of vocal fold function will not only provide additional insights into the biomechanics of normal and disordered voice production, but also has the potential to assist in planning phonosurgery by predicting the impact of a planned procedure on vocal function.

Conclusion

Within the past two years there have been published reports of new developments in auditory perceptual, acoustic, aerodynamic, and endoscopic imaging approaches for assessing voice quality and production. Some of the advances in perceptual (CAPE-V), acoustic (cepstral-based methods), and aerodynamic (phonation threshold pressure) assessment seem to have a better chance of being more rapidly adopted in the clinic, while others will require further development and testing.

References

1. American Speech-Language-Hearing Association. [Accessed December 19, 2007.] Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V). Available at: http://www.asha.org/about/membership-certification/divs/div_3.htm
2. Kreiman J, Gerratt BR. Sources of listener disagreement in voice quality assessment. *Journal of the Acoustical Society of America*. 2000; 108:1867–1876. [PubMed: 11051513]
- 3**. Kreiman J, Gerratt BR, Ito M. When and why listeners disagree in voice quality assessment tasks. *Journal of the Acoustical Society of America*. 2007; 122:2354–2364. This is the first study that quantifies the extent to which experimental design factors affect listener variability. Cf. [4]. [PubMed: 17902870]
- 4**. Shrivastav R, Sapienza CM, Nandur V. Application of psychometric theory to the measurement of voice quality using rating scales. *Journal of Speech, Language, and Hearing Research*. 2005; 48:323–335. This study espouses another approach to solve the listener reliability issue by applying psychometric theory. Cf. [3].
- 5**. Zeitels SM, Blitzer A, Hillman RE, et al. Foresight in laryngology and laryngeal surgery: A 2020 vision. *Annals of Otolaryngology, Rhinology, and Laryngology Supplement*. 2007; 198:2–16. This special supplement proposes future directions in laryngology that could be pursued and accomplished by the year 2020.
6. Awan SN, Roy N. Toward the development of an objective index of dysphonia severity: A four-factor acoustic model. *Clinical Linguistics and Phonetics*. 2006; 20:35–49. [PubMed: 16393797]
- 7*. Murphy PJ. On first harmonic amplitude in the analysis of synthesized aperiodic voice signals. *Journal of the Acoustical Society of America*. 2006; 120:2896–2907. This study is the first attempt to quantitatively explain what cepstral analysis reveals when analyzing voice samples. [PubMed: 17139747]
- 8*. Murphy PJ, Akande OO. Noise estimation in voice signals using short-term cepstral analysis. *Journal of the Acoustical Society of America*. 2007; 121:1679–1690. This study applies cepstral analysis to derive a noise-to-harmonics ratio measure for synthesized vowels with varying noise levels. [PubMed: 17407904]
- 9*. Zhang Y, Jiang JJ. Acoustic analyses of sustained and running voices from patients with laryngeal pathologies. *Journal of Voice*. 2008; 22:1–9. This article presents initial results on the use of nonlinear dynamics-based algorithms on running speech to better differentiate normal from disordered voices versus traditional perturbation analysis. [PubMed: 16978835]
- 10*. Chan RW, Titze IR. Dependence of phonation threshold pressure on vocal tract acoustics and vocal fold tissue mechanics. *Journal of the Acoustical Society of America*. 2006; 119:2351–2362. This study correlates phonation threshold pressure with vocal fold tissue properties in both a mathematical vocal fold model and a physical model of the larynx. [PubMed: 16642848]
- 11*. Jiang J, Leder C, Bichler A. Estimating subglottal pressure using incomplete airflow interruption. *Laryngoscope*. 2006; 116:89–92. This study introduces the partial air flow interruption method to estimate subglottal pressure, minimizing effects related to abrupt flow changes. [PubMed: 16481816]
- 12*. Baggott CD, Yuen AK, Hoffman MR, et al. Estimating subglottal pressure via airflow redirection. *Laryngoscope*. 2007; 117:1491–1495. This study introduces a system for estimating subglottal pressure that may be easily transferred to a clinical setting. [PubMed: 17762273]
- 13*. Jiang JJ, Tao C. The minimum glottal airflow to initiate vocal fold oscillation. *Journal of the Acoustical Society of America*. 2007; 121:2873–2881. This article introduces the “phonation threshold flow” parameter and demonstrates, using a mathematical model, the parameter’s sensitivity to simulated vocal fold properties. [PubMed: 17550186]
- 14**. Howe MS, McGowan RS. Sound generated by aerodynamic sources near a deformable body, with application to voiced speech. *Journal of Fluid Mechanics*. 2007; 592:367–392. This article summarizes the aeroacoustic theory of sound production and presents a detailed analysis of different source mechanisms that contribute to voice.
- 15*. McGowan RS, Howe MS. Compact Green’s functions extend the acoustic theory of speech production. *Journal of Phonetics*. 2006; 35:259–270. This article presents aeroacoustic

mathematical theory as an extension of the well-known linear source-filter theory of sound production.

- 16*. Deliyski DD, Petrushev PP, Bonilha HS, et al. Clinical implementation of laryngeal high-speed videoendoscopy: Challenges and evolution. *Folia Phoniatrica et Logopaedica*. 2007; 60:33–44. This article comprehensively presents results the latest results of the utility of high-speed imaging, as well as new facilitative visualizations that promise to offer clinically-salient information. [PubMed: 18057909]
- 17*. Shaw HS, Deliyski DD. Mucosal wave: A normophonic study across visualization techniques. *Journal of Voice*. 2008; 22:23–33. The goal of this study is to determine the variations in image-based judgments using stroboscopic and high-speed methods. [PubMed: 17014988]
- 18*. Švec JG, Šram F, Schutte HK. Videokymography in voice disorders: What to look for? *Annals of Otolaryngology, Rhinology, and Laryngology*. 2007; 116:172–180. This article presents a systematic methodology for categorizing kymographic data.
19. Yan Y, Chen X, Bless D. Automatic tracing of vocal-fold motion from high-speed digital images. *IEEE Transactions on Biomedical Engineering*. 2006; 53:1394–1400. [PubMed: 16830943]
- 20*. Yan YL, Damrose E, Bless D. Functional analysis of voice using simultaneous high-speed imaging and acoustic recordings. *Journal of Voice*. 2007; 21:604–616. This study exemplifies the image-analysis approach of deriving higher-order measures of vocal function from high-speed imaging methods. [PubMed: 16968665]
- 21*. Tao C, Zhang Y, Jiang JJ. Extracting physiologically relevant parameters of vocal folds from high-speed video image series. *IEEE Transactions on Biomedical Engineering*. 2007; 54:794–801. This study links asymmetric vocal fold model parameters with phonatory values derived from an excised preparation. [PubMed: 17518275]
- 22*. Schwarz R, Hoppe U, Schuster M, et al. Classification of unilateral vocal fold paralysis by endoscopic digital high-speed recordings and inversion of a biomechanical model. *IEEE Transactions on Biomedical Engineering*. 2006; 53:1099–1108. This study attempts to use an asymmetric vocal fold model to retroactively classify a particular voice disorder. [PubMed: 16761837]
- 23*. Wurzbacher T, Schwarz R, Döllinger M, et al. Model-based classification of nonstationary vocal fold vibrations. *Journal of the Acoustical Society of America*. 2006; 120:1012–1027. This study utilizes a mathematical model of the vocal folds that takes into account time-varying asymmetric vibratory patterns occurring during normal and disordered human phonation. [PubMed: 16938988]

Figure 1.

(a) Frames of mucosal wave (MW) playback (top) at different phases of the intra-glottal cycle and the corresponding 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). (b) A medial position frame of a mucosal wave kymography (MKG) playback. The MKG image brightness relates to the speed of motion of the mucosal edges, and the color shows the phase of motion (opening is green, closing is red). The mucosal wave extent appears as a double-edged or thicker curve during the closing phase. From Figures 3 and 4 in [16]. Used with permission.