Consensus Auditory-Perceptual Evaluation of Voice: Development of a Standardized Clinical Protocol

Gail B. Kempster
Rush University, Chicago

Bruce R. Gerratt
University of California, Los Angeles

Katherine Verdolini Abbott
University of Pittsburgh, Pittsburgh, PA

Julie Barkmeier-Kraemer
University of Arizona, Tucson

Robert E. Hillman
Massachusetts General Hospital, Boston

Purpose: This article presents the development of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) following a consensus conference on perceptual voice quality measurement sponsored by the American Speech-Language-Hearing Association's Special Interest Division 3, Voice and Voice Disorders. The CAPE-V protocol and recording form were designed to promote a standardized approach to evaluating and documenting auditory-perceptual judgments of vocal quality.

Method: A summary of the consensus conference proceedings and the factors considered by the authors in developing this instrument are included.
Conclusion: The CAPE-V form and instructions, included as appendices to this article, enable clinicians to document perceived voice quality deviations following a standard (i.e., consistent and specified) protocol.

Key Words: Consensus Auditory-Perceptual Evaluation of Voice, voice, voice assessment

The Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) is a clinical and research tool developed to promote a standardized approach to evaluating and documenting auditory-perceptual judgments of voice quality. The tool was created as a direct outcome of the Consensus Conference on Auditory-Perceptual Evaluation of Voice, held in June 2002 and sponsored by the American Speech-Language-Hearing Association's (ASHA) Special Interest Division 3, Voice and Voice Disorders and the University of Pittsburgh. The purpose of this article is to document the development of the CAPE-V protocol and form, and provide a rationale for each of the elements included in the protocol.

The consensus conference brought together an international group of voice scientists, experts in human perception, and speech-language pathologists to explore solutions to a long-standing need in clinical voice pathology: to apply scientific evidence about psychophysical measurement to the clinical practice of judging auditory-perceptual features of voice quality. (See Appendix A for a list of conference participants.) Following 2 days of presentations and discussion, recommendations from these participants informed and guided the development of the CAPE-V tool. The CAPE-V authors (the authors of this article) approached the task of psychophysical measurement and the scaling of voice quality by adhering to the consensus opinions of scientists and clinicians. From its inception, the CAPE-V was intended to become a standardized protocol, useful to clinicians and researchers, that incorporates multiple recommendations for best practices in assessing perceived abnormal vocal quality (Barkmeier, Verdolini, & Kempster, 2002). The word "standardized" is used throughout this article to refer to a

procedure that is administered and scored in a consistent way; it does not here denote norm-referencing.

The continuum of normal to abnormal voice quality is inextricably related to vocal health. While auditory-perceptual judgments of voice and speech can never be accomplished with perfect validity or reliability (Gerratt & Kreiman, 2000; Gerratt, Kreiman, Antonanzas-Barroso, & Berke, 1993; Kent, 1996; Kreiman & Gerratt, 1998, 2000; Kreiman, Gerratt, Kempster, Erman, & Berke, 1993; Kreiman, Gerratt, Precoda, & Berke, 1992; Shrivastav, 2006; Shrivastav, Sapienza, & Nandur, 2005), perceptual appraisal of voice quality remains a key standard for judgment of vocal impairments, both for patients who experience vocal changes and for the clinicians who treat them (Carding, Carlson, Epstein, Mathieson, & Shewell, 2000; Hirano, 1981; Oates & Russell, 1998; Wilson, 1987). Simply stated, auditory-perceptual measures of voice quality define the presence or absence of a voice disorder clinically. Voice clinicians who treat these patients make auditory-perceptual judgments. Thus, there is a clear need for a way to make such judgments that is sound theoretically, is clinically meaningful, and can be consistently administered.

Consensus Conference Issues and Summary

The 2-day conference began with a statement of the problem, that of creating valid and reliable measures of auditory-perceptual features of voice quality. Four invited scientists explored issues surrounding the difficult task of psychophysical measurement and scaling, as understood from relevant areas of human perception. The presentations reviewed the historical background of auditory-perceptual evaluation of voice and speech and described the state of the art in human auditory perception, with particular emphasis on how such information might affect the auditory-perceptual assessment of voice disorders. Several voice researchers, from the United States and elsewhere, added information from their investigations of voice quality. Finally, a report on routine clinical practice in the United States was included to relate current practice patterns to conclusions drawn from the scientific discussion. Throughout the conference, the scientists and clinicians reacted to the clinical and research conundrums in auditory-perceptual judgments of voice quality and the challenge of developing a new assessment instrument. At the conclusion of the conference, the authors of this article collaborated to draft the CAPE-V form and procedures.

Auditory-Perceptual Evaluation: Exploring the Elusive Ideal

Raymond Kent provided a broad review of auditory-perceptual assessment of voice and speech and outlined the challenges and assumptions associated with establishing an "ideal" perceptual evaluation method. In voice assessment, such an "ideal" method would (a) provide a reliable means of differentiating normal and disordered voices, and tracking changes in a patient's vocal status across time; (b) correlate with underlying pathophysiology and objective measures; and (c) be clearly established, including type of

scale(s), type and amount of user training needed, and whether to use anchors in training. The establishment of an "ideal" method also requires that some well-known obstacles be overcome. These include the lack of standard terminology for describing or scaling disordered voice quality, the absence of a standard definition of "normal voice," inherent poor reliability of auditory-perceptual judgments of voice quality, and inherent variability of an individual's voice production. Although Kent conceded that the perceptual assessment of voice quality has been "an uncertain endeavor, vexed by disagreements among authorities and variability in data," he cited reasons for some optimism, based on a growing international interest in developing a standardized procedure, as evidenced by the international representation at the Consensus Conference. He also pointed to emerging consensus points on some basic issues, including what kind of scale to use and how many and which attributes to rate. Moreover, recent demonstrations suggest that computer modeling and interactive synthesis of disordered voice quality can assist in developing improved methods for auditory-perceptual assessment (Callan, Kent, Roy, & Tasko, 2000; Gerratt & Kreiman, 2001; Kreiman & Gerratt, 2000).

Psychoacoustic Principles and Human Perception

Lawrence Feth reviewed current psychoacoustics-based perspectives on human perception and the discrimination of sound (Houtsma, 1995; Zwicker, Fastl, & Frater, 1999). He presented a brief review of the anatomy and physiology of the ear, as well as an overview of how sound is processed going from peripheral to central auditory mechanisms. He described (a) the peripheral influences of the outer and middle ear as manifested by the audibility curve, (b) the frequency selectivity/critical bandwidth processing and intensity compression that are initially the result of cochlear morphology and biomechanics, and (c) integration of acoustic information (e.g., spectral integration) that takes place at higher levels of the central nervous system. Feth also summarized what is currently known about how humans perceive and discriminate acoustic parameters of sound. Most of the work in this area has focused on pitch and loudness perception, which has influenced the scaling methods and theoretical constructs for both of these phenomena. Much less effort has been expended in formally studying the perception of sound *quality,* primarily because it is a more complex (multidimensional) and difficult to quantify perceptual phenomenon. Two sounds that are judged to have equal pitch and loudness but can still be discriminated from each other are said to differ in *timbre* or *quality.* By way of example, Feth briefly summarized some of the work by Zwicker et al. (1999) in which they attempted to explore the perception of quality-related concepts such as *sharpness, pleasantness, fluctuation, strength,* and *roughness.*

Psychophysical Issues Related to Scaling

George Gescheider and Lawrence Marks each gave presentations dealing with the psychophysical bases of perceptual scaling and measurement (Gescheider & Marks, 2002; Marks & Algom, 1998). Gescheider briefly reviewed

Kempster et al.: *Consensus Auditory-Perceptual Evaluation of Voice* **125**

the classic work of some well-known pioneers in psycho-physics (Fechner, Weber, Stevens, and Thurstone) in discussing basic approaches for determining absolute and difference (just noticeable differences or difference limens) thresholds for sensory systems. He concluded with an overview of additional scaling methods, including partition, ratio, and multidimensional approaches.

Marks addressed the methodological issues in perceptual scaling and measurement more specifically, explaining that the general process of psychophysical analysis involves (a) definition of the stimuli (*What properties of stimuli are pertinent to perception?*), (b) definition of the perceptual experiences (*What are the attributes/features/dimensions of perception and how do these attributes/features/dimensions interrelate?*), and (c) determination/modeling of the processes that relate percepts to stimuli (*sensory, decisional, cognitive processes: encoding, transforming, recoding, etc.*).

Marks also commented on the concepts of internal and external validity, sensitivity, and reliability of scaling procedures, as well as differences between direct as compared with indirect scaling methods. Direct methods, which are considered more appropriate than indirect approaches for clinical applications, make use of interval or ratio scaling procedures. To optimize direct scaling, consideration should be given to the actual range and distribution of the stimuli being used, whether standard or anchor stimuli are employed, whether there are sequential/order effects, and whether training improves performance (Gescheider & Marks, 2002; Marks & Algom, 1998).

Marks made two specific recommendations with respect to developing a clinical instrument for auditory-perceptual assessment of voice quality. First, he recommended using numerical rating scales with at least 15 subcategories/ divisions or, alternatively, employing continuous graphical/ visual analog scales. Second, he recommended that the location of anchors (e.g., *normal* or *most severe*) be adjusted to provide extra room at the ends of the scale to avoid end effects.

Current Practice

Reports from international experts present at the conference reviewed the utilization of auditory-perceptual scales in clinical practice and research, including interactive training models and the use of training scales and anchors (Chan & Yiu, 2002; Oates & Russell, 1998) and other formalized perceptual scaling instruments and procedures including Vocal Profile Analysis by Laver (Carding et al., 2000) and the Stockholm Voice Evaluation Approach (Hammarberg, 2000). Carding reviewed current methods in Britain and noted that while most clinicians in the United Kingdom are trained in Laver's Vocal Profile Analysis, the GRBAS (grade, rough, breathy, asthenic, and strained) method (Hirano, 1981) is recommended as the minimum standard for practicing voice clinicians in the United Kingdom. Several participants referred to other influential sources of information related to the perception of vocal quality: Kreiman et al. (1993) and DeBodt, Wuyts, Van de Heyning, and Croux (1997). Kreiman et al. identified 57 different perceptual

schemes for voice assessment and concluded that the most widely used was the Buffalo Voice Profile (Wilson, 1987). Work in Belgium by DeBodt and his colleagues includes clinical recommendations about appropriate use of various options based on a review of contemporary perceptual rating scales.

Consensus Points

The conference attendees agreed that there is no single, best way to approach the task of measuring perceived vocal quality. The current knowledge base is inadequate for designing a clinical tool that resolves all of the relevant scientific issues. Indeed, efforts to do so have reflected an array of problems of reliability, utility, and validity, and these limitations are also true in the development of the CAPE-V. Nonetheless, the CAPE-V authors incorporated multiple perspectives, from scientific data to clinical practice, to develop both a protocol to follow and a form to document auditory-perceptual features of abnormal quality. Conference participants agreed that constructing a consistent and specified set of evaluation procedures and a documentation format would, at a minimum, improve communication and consistency among clinicians.

In this context, the authors agreed on the following orienting principles:

1. Perceptual dimensions should reflect a minimal set of clinically meaningful, perceptual voice parameters.
2. Procedures and results should be obtainable expediently.
3. Procedures and results should be applicable to a broad range of vocal pathologies and clinical settings.
4. Ratings should be demonstrated to optimize reliability within and across clinicians through later validation studies.
5. Ultimately, exemplars may be considered for future use as anchors and possibly for training.

Specific Elements of the CAPE-V

The CAPE-V instructions are included as Appendix B, and the form for documenting the assessment is presented in Appendix C.

Tasks

The CAPE-V stipulates that the individual whose voice is to be assessed (hereafter referred to as the "patient") perform three specific vocal tasks: (a) sustain the vowels /a/ and /i/ three times each; (b) read six specific sentences with different phonetic contexts; and (c) converse naturally in response to the standard question ("Tell me about your voice problem").

Rationale for the tasks. The first task elicits vowel prolongations. Vowel prolongations (at a steady and comfortable pitch level) provide an opportunity to listen to a patient's voice without articulatory influences. Vowels can also be analyzed acoustically, for which some normative data are available. The second task elicits six sentences of varied

speech contexts from which to assess different elements of vocal quality. Sentence 1 ("The blue spot is on the key again") is a commonly used stimulus sentence to examine the coarticulatory influence of three vowels (/a, i, u/). Sentence 2 ("How hard did he hit him?") provides a context to assess soft glottal attacks and voiceless to voiced transitions. Sentence 3 ("We were away a year ago") features all voiced phonemes and provides a context to judge possible voiced stoppages/spasms and one's ability to "link" (i.e., maintain voicing) from one word to another. Sentence 4 ("We eat eggs every Easter") includes several vowel-initiated words that may provoke hard glottal attacks and provides the opportunity to assess whether these occur. Sentence 5 ("My mama makes lemon jam") includes numerous nasal consonants, thus providing an opportunity to assess hyponasality and possible stimulability for resonant voice therapy. Finally, Sentence 6 ("Peter will keep at the peak") contains no nasal consonants and provides a useful context for assessing intraoral pressure and possible hypernasality or nasal air emission. The third task elicits conversational speech and is the most important and relevant to both patient and clinician. Although, in the CAPE-V protocol, conversation is assessed after the vowels and sentences, it is expected that this aspect of the patient's voice is under close observation throughout the evaluation session.

Quality Features to be Assessed

The CAPE-V protocol specifies six quality features to be evaluated consistently and allows flexibility to add other perceptual features of interest. The six voice quality features selected for consistent appraisal are labeled and defined as follows:

Overall Severity: global, integrated impression of voice deviance

Roughness: perceived irregularity in the voicing source

Breathiness: audible air escape in the voice

Strain: perception of excessive vocal effort (hyperfunction)

Pitch: perceptual correlate of fundamental frequency

Loudness: perceptual correlate of sound intensity

Rationale for the quality features. Despite much debate over the description, validity, and independence of any list of voice quality features, these six have consistently appeared in both national and international voice literature for decades (DeBodt et al., 1997; Fairbanks, 1960; Hirano, 1981; Wilson, 1987). Thus, the rationale for including these six voice quality features is the belief that both clinicians and researchers find these attributes meaningful. Another common descriptor, *hoarse,* was excluded from the list of terms because the authors agreed with Fairbanks (1960) that "hoarseness" is perceived by many as a combination of "roughness" and "breathiness." The CAPE-V form also includes two unlabeled scales. These allow the clinician to document other salient perceptual features of a patient's voice, such as degree of nasality, spasm, tremor, intermittent aphonia, falsetto, glottal fry, or weakness.

Scale

A 100-mm line scale with unlabeled anchors, commonly known as a visual analog scale, is used to assess each of the six quality features. The leftmost portion of the scale reflects normal voice (in the case of judging severity, pitch, or loudness) or none of the quality being judged (in the case of roughness, breathiness, and strain). The right end of the scale is to reflect the listener's judgment of the most extreme example of deviance. A tick mark for each of the three tasks, with the subscript 1 (for vowels), 2 (for sentences), and/or 3 (for conversation) is drawn onto the scale to reflect a listener's judgment for each scale. Measurement from the left end of the scale to each tick mark, in millimeters, is denoted on the blank to the far right of the scale (___/100).

Rationale for the scale. Marks recommends that auditory-perceptual judgments of voice quality be made on a visual analog scale (or set of scales), using open-ended anchor points at either end as a way to inhibit end effects of the scale. Visual analog scales are easy for raters to use and appear to have become more commonplace in voice research in the past 2 decades.

Verbal Descriptor Degree of Deviance

While the primary measurement index is an interval scale provided by the 100-mm visual analog line, the CAPE-V also includes the ordinal ratings of *mild, moderate,* and *severe,* printed below the measurement line, to serve as a supplemental severity indicator. These qualitative terms are positioned in a nonequidistant fashion, based on Marks's recommendations, and reflect the range of voice severity using terminology more familiar to clinicians than the discrete intervals measured on the 100-mm visual analog scale.

Additional CAPE-V Elements

A nominal rating judgment allows the clinician to classify the consistency or intermittent presence of the voice quality feature within and across evaluation tasks. Sections devoted to resonance or other features supplement the CAPE-V protocol by allowing other salient descriptors to document a patient's voice quality. This flexibility is needed to capture the spectrum of voice disorders and associated conditions or features. The list of terms provided on the form is not inclusive, meant only as examples of specific features that may help describe auditory-perceptual attributes.

Rating Procedures

The CAPE-V judgments are intended to reflect the clinician's direct observations of the patient's performance during the evaluation and should not take into account patient report or other sources. Standard audio-recording procedures should be used, such as recording in a quiet environment and using a standard mouth-to-microphone distance with the highest possible sampling rate for digital conversion. If a patient returns following an initial assessment, the clinician may compare the initial voice sample and

CAPE-V ratings directly to any subsequent recordings, to optimize the internal consistency or reliability of repeated sequential ratings, particularly for assessing treatment outcomes. As always, clinicians are encouraged to minimize bias in all ratings.

The CAPE-V form and instructions have been available to affiliates of Special Interest Division 3 on the protected portion of the division's Web site since 2003. The tool was presented at national conventions as early as 2002 (Barkmeier, 2003; Barkmeier et al., 2002; Hillman, 2003; Shrivastav, Kempster, & Zraick, 2006; Zraick et al., 2007). The instrument is already used in more than 20 clinics and some laboratories throughout the United States, and using the CAPE-V protocol as directed has been shown to add no more than a few minutes to a voice evaluation session (M. Spencer, personal communication, June 16, 2005). A national, multi-institutional validation study examining the reliability of the instrument has also begun (Zraick et al., 2007).

Concurrent Validity and the CAPE-V

A master's thesis (Berg & Eden, 2003) directed by Hammarberg and Holmberg compared aspects of the CAPE-V to the Stockholm Voice Evaluation Approach on patients with three different voice pathologies (E. Holmberg, personal communication, December 1, 2003). This study involved a translation of the CAPE-V into Swedish. The authors determined that intra- and interrater reliability was acceptably high in both protocols, and no obvious differences were found between the two approaches in terms of listener variability. Both protocols were able to separate the three disorders from each other and showed significant pre-to-posttreatment changes in voice quality.

Karnell et al. (2007) published a preliminary report comparing the reliability of clinician-based auditory-perceptual judgments using the CAPE-V to those made with the GRBAS voice-rating scheme (Hirano, 1981) and two other quality of life scales. Among other findings, Karnell et al. found comparable estimates of interrater reliability for the two scales, both at high levels. They suggest that the CAPE-V may offer "more sensitivity to small differences within and among patients than the GRBAS scale" (p. 1).

A second preliminary investigation has suggested that the CAPE-V results meet or exceed the GRBAS in measurement reliability (Zraick et al., 2007).

The CAPE-V's similarity to the GRBAS scale is obvious to anyone familiar with both scales. In fact, the CAPE-V uses all of the GRBAS percepts (except aesthenic) for judging voice quality, and the definitions of the quality features are also similar. However, three important factors discriminate the CAPE-V from the GRBAS scale. First, the GRBAS has no published, standardized protocol to follow in English. The Hirano (1981) reference most often cited for the GRBAS provides no guidelines for clinical administration, speech material, or rating calibration. In contrast, the CAPE-V includes a specific protocol that designates the tasks, procedures, and scaling routine, toward the larger goal of improving the consistency of clinical assessment from one clinician to another, without excessive demands on clinician time or learning. Second, the CAPE-V provides interval

scale measures of voice quality by incorporating millimeter measures on visual analog scales. Such scales are shown to better accommodate the task of measurement of multidimensional features, such as vocal quality (Chan & Yiu, 2002; Gerratt et al., 1993). The GRBAS scale, however, only allows ordinal judgments on a four-point scale of normal (1), mild (2), moderate (3), or severe (4), which severely limits its application to research design and statistical analysis. Finally, the CAPE-V attempts to document more voice quality features than the GRBAS, across more speech tasks, while allowing room for supplemental feature scales and comment areas.

Summary and Conclusions

The CAPE-V is the result of an effort sponsored by ASHA Special Interest Division 3 to create a clinical protocol that can be used for making auditory-perceptual judgments of voice quality in a standardized way. The CAPE-V form and procedures represent the consensus recommendations from experts in human perception, speech and voice scientists, and speech-language pathologists who specialize in voice disorders. Although there is no known ideal method for obtaining reliable and valid judgments of auditory-perceptual features, the CAPE-V derives its protocol and measurement scales from a state-of-the-art understanding of the multidimensional factors that underlie psychophysical measurement and human perception. As such, the authors hope it serves to support and encourage best clinical practices in the auditory-perceptual evaluation of voice.

Acknowledgments

References

Barkmeier, J. (2003, November). *Update on the use of the CAPE-V tool*. Presented at the Annual Convention of the American Speech-Language-Hearing Association, Chicago.

Barkmeier, J., Verdolini, K., & Kempster, G. (2002, November). *Report of the consensus conference on auditory-perceptual evaluation of voice*. Presented at the Annual Convention of the American Speech-Language-Hearing Association, Atlanta, GA.

Berg, B., & Edén, S. (2003). Perceptuell bedömning av röstkvalitet hos tre organiska röststörningar—jämförelse mellan Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) och Stockholm Voice Evaluation Approach (SVEA). [Perceptual evaluation of

voice quality in three organic voice disorders—a comparison between Consensus Auditory Perceptual Evaluation of Voice (CAPE-V) and Stockholm Voice Evaluation Approach (SVEA)]. Unpublished master's thesis, Karolinska Institute, Stockholm, Sweden.

Callan, D., Kent, R., Roy, N., & Tasko, S. (2000). The use of self-organizing maps for the classification of voice disorders. In R. Kent & M. Ball (Eds.), *Voice quality measurement* (pp. 103–116). San Diego, CA: Singular.

Carding, P. N., Carlson, E., Epstein, R., Mathieson, L., & Shewell, C. (2000). Formal perceptual evaluation of voice quality in the United Kingdom. *Logopedics, Phoniatrics, Vocology, 25*(3), 133–138.

Chan, K., & Yiu, E. (2002). The effect of anchors and training on the reliability of perceptual voice evaluation. *Journal of Speech, Language, and Hearing Research, 45,* 111–126.

DeBodt, M. S., Wuyts, F. L., Van de Heyning, P. H., & Croux, C. (1997). Test-retest study of the GRBAS scale: Influence of experience and professional background on perceptual rating of voice quality. *Journal of Voice, 11*(1), 74–80.

Fairbanks, G. (1960). *Voice and articulation drillbook.* Philadelphia: Harper & Row.

Gerratt, B. R., & Kreiman, J. (2000). Theoretical and methodological development in the study of pathological voice quality. *Journal of Phonetics, 28*(3), 335–342.

Gerratt, B. R., & Kreiman, J. (2001). Measuring vocal quality with speech synthesis. *Journal of the Acoustical Society of America, 110,* 2560–2566.

Gerratt, B., Kreiman, J., Antonanzas-Barroso, N., & Berke, G. (1993). Comparing internal and external standards in voice quality judgments. *Journal of Speech and Hearing Research, 36,* 14–20.

Gescheider, G. A., & Marks, L. E. (2002). Psychophysical scaling. In H. E. Pashler & S. S. Stevens (Eds.), *Stevens' handbook of experimental psychology* (3rd ed., pp. 91–138). Indianapolis, IN: Wiley.

Hammarberg, B. (2000). Voice research and clinical needs. *Folia Phoniatrica et Logopaedica, 52,* 93–102.

Hillman, R. (2003, June). *Overview of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) instrument developed by ASHA Special Interest Division 3.* Presented at the Annual Symposium on the Care of the Professional Voice, Philadelphia.

Hirano, M. (1981). *Clinical examination of voice.* New York: Springer Verlag.

Houtsma, A. J. M. (1995). Pitch perception. In B. C. J. Moore (Ed.), *Handbook of perception and cognition* (2nd ed., pp. 276–295). San Diego, CA: Academic Press.

Karnell, M., Melton, S., Childes, J., Coleman, T., Dailey, S., & Hoffman, H. (2007). Reliability of clinician-based (GRBAS and CAPE-V) and patient-based (V-RQOL and IPVI) documentation of voice disorders. *Journal of Voice, 21,* 576–590.

Kent, R. D. (1996). Hearing and believing: Some limits to the auditory-perceptual assessment of speech and voice disorders. *American Journal of Speech-Language Pathology, 5*(3), 7–23.

Kreiman, J., & Gerratt, B. R. (1998). Validity of rating scale measures of voice quality. *Journal of the Acoustical Society of America, 104,* 1598–1608.

Kreiman, J., & Gerratt, B. R. (2000). Sources of listener disagreement in voice quality assessment. *Journal of the Acoustical Society of America, 108,* 1867–1876.

Kreiman, J., Gerratt, B. R., Kempster, G. B., Erman, A., & Berke, G. S. (1993). Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research. *Journal of Speech, Language, and Hearing Research, 36,* 21–40.

Kreiman, J., Gerratt, B. R., Precoda, K., & Berke, G. (1992). Individual differences in voice quality perception. *Journal of Speech and Hearing Research, 35,* 512–520.

Marks, L., & Algom, D. (1998). Psychophysical scaling. In M. H. Birnbaum (Ed.), *Measurement, judgment, and decision making.* San Diego, CA: Academic Press.

Oates, J., & Russell, A. (1998). Learning voice analysis using an interactive multi-media package: Development and preliminary evaluation. *Journal of Voice, 12,* 500–512.

Shrivastav, R. (2006). Multidimensional scaling of breathy voice quality: Individual differences in perception. *Journal of Voice, 20,* 211–222.

Shrivastav, R., Kempster, G., & Zraick, R. (2006, November). *Now hear this: Improving perceptual evaluations of voice quality.* Paper presented at the Annual Convention of the American Speech-Language-Hearing Association, Miami Beach, FL.

Shrivastav, R., Sapienza, C. M., & Nandur, V. (2005). Application of psychometric theory to the measurement of voice quality using rating scales. *Journal of Speech, Language, and Hearing Research, 48,* 323–335.

Wilson, D. K. (1987). *Voice problems of children* (3rd ed.). Baltimore: Williams and Wilkins.

Zraick, R., Klaben, B., Connor, N., Thibeault, S., Glaze, L., Thrush, C., & Bursac, Z. (2007, November). *Results of the CAPE-V validation study.* Paper presented at the Annual Convention of the American Speech-Language-Hearing Association, Boston.

Zwicker, E., Fastl, H., & Frater, H. (1999). *Psychoacoustics: Facts and models.* New York: Springer-Verlag.

Contact author: Gail B. Kempster, Department of Communication Disorders and Sciences, Rush University Medical Center, 1653 W. Congress Parkway, 203 Senn, Chicago, IL 60612. E-mail: gail_b_kempster@rush.edu.

Appendix A

Participants at the Consensus Conference on Auditory-Perceptual Assessment of Voice Disorders, University of Pittsburgh, June 10–11, 2002

Julie Barkmeier	Rebecca Leonard
Diane Bless	Lori Lombard
Paul Carding	Christy Ludlow
Karen Chan	Lawrence Marks
Raymond Colton	Malcolm McNeil
Mary Erickson	Thomas Murry
Michelle Ferketic	Jennifer Oates
Lawrence Feth	Kristin Pelczarski
Bruce Gerratt	Lorraine Ramig
George Gescheider	Doug Roth
Leslie Glaze	Mary Sandage
Douglas Hicks	Christine Sapienza
James Hillenbrand	Lana Shekim
Robert Hillman	Rahul Shrivastav
Eva Holmberg	Kim Steinhauer
Celia Hooper	Joseph Stemple
Michael Karnell	Johann Sundberg
Gail Kempster	Micheal Trudeau
Raymond Kent	Katherine Verdolini
Jody Kreiman	Edwin M-L. Yiu

The authors regret if they have inadvertently omitted a participant from this list.

Appendix B (p. 1 of 2)

CAPE-V Procedures

Developed by Gail Kempster, Bruce Gerratt, Katherine Verdolini, Julie-Barkmeier-Kraemer, and Robert Hillman (June 2002)
ASHA Special Interest Division 3, Voice and Voice Disorders

Description and Instructions

General description of the tool. The CAPE-V indicates salient perceptual vocal attributes, identified by the core consensus group as commonly used and easily understood. The attributes are: (a) Overall Severity; (b) Roughness; (c) Breathiness; (d) Strain; (e) Pitch; and (f) Loudness. The CAPE-V displays each attribute accompanied by a 100-mm line forming a visual analog scale (VAS). The clinician indicates the degree of perceived deviance from normal for each parameter on this scale, using a tick mark. For each dimension, scalar extremes are unlabeled. Judgments may be assisted by referring to general regions indicated below each scale on the CAPE-V: "MI" refers to "mildly deviant," "MO" refers to "moderately deviant," and "SE" refers to "severely deviant." A key issue is that the regions indicate gradations in severity, rather than discrete points. The clinician may place tick marks at any location along the line. Ratings are based on the clinician's direct observations of the patient's performance during the evaluation, rather than patient report or other sources.

To the right of each scale are two letters, C and I. C represents "consistent," and I represents "intermittent" presence of a particular voice attribute. The rater circles the letter that best describes the consistency of the judged parameter. A judgment of "consistent" indicates that the attribute was continuously present throughout the tasks. A judgment of "intermittent" indicates that the attribute occurred inconsistently within or across tasks. For example, an individual may consistently exhibit a strained voice quality across all tasks, which include sustained vowels and speech. In this case, the rater would circle C to the right of the strain scale. In contrast, another individual might exhibit consistent strain during vowel production, but intermittent strain during one or more connected speech task. In this case, the rater would circle I to the right of the strain scale.

Definitions of vocal attributes. The features of voice that are to be rated are defined as follows:

- **Overall Severity:** global, integrated impression of voice deviance
- **Roughness:** perceived irregularity in the voicing source
- **Breathiness:** audible air escape in the voice
- **Strain:** perception of excessive vocal effort (hyperfunction)
- **Pitch:** perceptual correlate of fundamental frequency. This scale rates whether the individual's pitch deviates from normal for that person's gender, age, and referent culture. The direction of deviance (high or low) should be indicated in the blank provided above the scale.
- **Loudness:** perceptual correlate of sound intensity. This scale indicates whether the individual's loudness deviates from normal for that person's gender, age, and referent culture. The direction of deviance (soft or loud) should be indicated in the blank provided above the scale.

Blank scales and additional features. The six standard vocal attributes included on the CAPE-V are considered the minimal set of parameters for describing the auditory-perceptual characteristics of disordered voices. The form also includes two unlabeled scales. The clinician may use these to rate additional prominent attributes required to describe a given voice. The clinician may indicate the presence of other attributes or "positive signs" not noted elsewhere under "Additional features." If an individual is aphonic, this should be noted under "Additional features," and no additional marks should be made on the scales.

CAPE-V Procedures

Data Collection

The individual should be seated comfortably in a quiet environment. The clinician should audio-record the individual's performance on three tasks: vowels, sentences, and conversational speech. Standard recording procedures should be used that incorporate a condenser microphone placed at an azimuth of 45° from the front of the mouth and at a 4-cm microphone-to-mouth distance. Audio recordings are recommended to be made onto a computer with 16 bits of resolution and a signal-sampling rate of no less than 20 KHz.

Task 1: Sustained vowels. Two vowels were selected for this task. One is considered a lax vowel (/a/) and the other tense (/i/). In addition, the vowel, /i/, is the sustained vowel used during videostroboscopy. Thus, the use of this vowel during this task offers an auditory comparison to that produced during a stroboscopic exam.

The clinician should say to the individual, "The first task is to say the sound, /a/. Hold it as steady as you can, in your typical voice, until I ask you to stop." (The clinician may provide a model of this task, if necessary.) The individual performs this task three times for 3 to 5 s each. "Next, say the sound, /i/. Hold it as steady as you can, in your typical voice, until I ask you to stop." The individual performs this task three times for 3 to 5 s each.

Task 2: Sentences. Six sentences were designed to elicit various laryngeal behaviors and clinical signs. The first sentence provides production of every vowel sound in the English language, the second sentence emphasizes easy onset with the /h/, the third sentence is all voiced, the fourth sentence elicits hard glottal attack, the fifth sentence incorporates nasal sounds, and the final sentence is weighted with voiceless plosive sounds.

The clinician should give the person being evaluated flash cards, which progressively show the target sentences (see below) one at a time. The clinician says, "Please read the following sentences one at a time, as if you were speaking to somebody in a real conversation." (Individual performs task, producing one exemplar of each sentence.) If the individual has difficulty reading, the clinician may ask him or her to repeat sentences after verbal examples. This should be noted on the CAPE-V form. The sentences are: (a) The blue spot is on the key again; (b) How hard did he hit him? (c) We were away a year ago; (d) We eat eggs every Easter; (e) My mama makes lemon jam; and (f) Peter will keep at the peak.

Task 3: Running speech. The clinician should elicit at least 20 s of natural conversational speech using standard interview questions such as "Tell me about your voice problem" or "Tell me how your voice is functioning."

Data Scoring

Although the PDF scale is accurate, printer configurations vary. Please verify that your paper copy has accurate 100-mm lines before reproducing the CAPE-V form (Appendix C). The clinician should have the individual perform all voice tasks—including vowel prolongation, sentence production, and running speech, before completing the CAPE-V form. If performance is uniform across all tasks, the clinician should mark the ratings indicating overall performance for each scale. If the clinician notes a discrepancy in performance across tasks, he or she should rate performance on each task separately, on a given line. Only one CAPE-V form is used per individual being evaluated. In the case of discrepancies across tasks, tick marks should be labeled with the task number. Tick marks reflecting vowel prolongation should be labeled #1 (see form). Tick marks reflecting running speech should be labeled #2. Tick marks reflecting story retelling should be labeled #3. In the rare event that the clinician perceives discrepancies within task type (e.g., /a/ vs. /i/), he or she may further label the ratings accordingly, such as 1/a/ versus 1/i/ to reflect the different vowels, or 2(a)-(b)-(c)-(d)-(e)- or (f) for the different sentences. Unlabeled tick marks indicate uniform performance. See examples below. (Note: Using labels to indicate discrepancies/variation across tasks in the severity of an attribute is different than indicating that an attribute is displayed intermittently [I]. If an attribute is judged to have equal severity whenever it appears, but it is not present all the time, "I" should be circled to indicate that the attribute is intermittent, and no additional labeling needs to be done.)

After the clinician has completed all ratings, he or she should physically measure the distance in millimeters from each scale. To do so, he or she should physically measure the distance in millimeters from the left end of the scale. The millimeters score should be written in the blank space to the far right of the scale, thereby relating the results in a proportion to the total 100-mm length of the line. The results can be reported in two possible ways. First, results can indicate distance in millimeters to describe the degree of deviancy, for example "73/100" on "strain." Second, results can be reported using descriptive labels that are typically employed clinically to indicate the general amount of deviancy, for example "moderate-to-severe" on "strain." We strongly suggest using both forms of reporting.

It is strongly recommended that for all rating sessions following the initial one, the clinician have a paper or electronic copy of the previous CAPE-V ratings available for comparison purposes. He or she should also rate subsequent examinations based on direct comparisons between earlier and current audio recordings. Such an approach should optimize the internal consistency/reliability of repeated sequential ratings within a patient, particularly for purposes of assessing treatment outcomes. Although difficult, clinicians are encouraged to make every effort to minimize bias in all ratings. We acknowledge that this solution is imperfect.

Other Elements

The clinician can indicate prominent observations about resonance phenomena under "Comments about resonance." Examples include, but are not limited to, hyper- or hyponasality and cul-de-sac resonance.

Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V)

Name:_____   **Date:**_____

The following parameters of voice quality will be rated upon completion of the following tasks:
1. Sustained vowels, /a/ and /i/ for 3-5 seconds duration each.
2. Sentence production:
      a.   The blue spot is on the key again.   d. We eat eggs every Easter.
      b.   How hard did he hit him?   e. My mama makes lemon muffins.
      c.   We were away a year ago.   f. Peter will keep at the peak.
3. Spontaneous speech in response to: "Tell me about your voice problem." or "Tell me how your voice is functioning."

Legend: C = Consistent   I = Intermittent
                MI = Mildly Deviant   MO = Moderately Deviant   SE = Severely Deviant
Although the PDF scale is accurate, printer configurations vary. Verify that your paper copy has accurate 100-mm lines before reproducing this form.

Overall Severity _____   C   I   ___/100
                      MI   MO   SE

Roughness   _____   C   I   ___/100
                      MI   MO   SE

Breathiness   _____   C   I   ___/100
                      MI   MO   SE

Strain   _____   C   I   ___/100
                      MI   MO   SE

Pitch   (Indicate the nature of the abnormality): _____
            _____   C   I   ___/100
                      MI   MO   SE

Loudness   (Indicate the nature of the abnormality): _____
            _____   C   I   ___/100
                      MI   MO   SE

_____   _____   C   I   ___/100
                      MI   MO   SE

_____   _____   C   I   ___/100
                      MI   MO   SE

COMMENTS ABOUT RESONANCE: NORMAL   OTHER (Provide description):_____
_____

ADDITIONAL FEATURES (for example, diplophonia, fry, falsetto, asthenia, aphonia, pitch instability, tremor, wet/gurgly, or other relevant terms):

                                                 Clinician:_____

Note. This form may be photocopied for clinical purposes. Available online at http://ajslp.asha.org.

Consensus Auditory-Perceptual Evaluation of Voice: Development of a Standardized Clinical Protocol

Gail B. Kempster, Bruce R. Gerratt, Katherine Verdolini Abbott, Julie Barkmeier-Kraemer, and Robert E. Hillman

AMERICAN
SPEECH-LANGUAGE-
HEARING
ASSOCIATION