Perspectives of the ASHA Special Interest Groups
SIG 3, Vol. 1(Part 2), 2016, Copyright © 2016 American Speech-Language-Hearing Association

Emerging Scientist: Challenges to CAPE-V as a Standard

Kathleen F. Nagle

Center for Laryngeal Surgery & Voice Rehabilitation, Massachusetts General Hospital
Boston, MA
Department of Speech-Language Pathology, Seton Hall University
South Orange, NJ

Disclosures

Kathleen F. Nagle is an Assistant Professor of Speech-Language Pathology at Seton Hall University
Nonfinancial: Kathleen F. Nagle has previously published in this topic area.

Abstract

The Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V; American Speech-Language-Hearing Association, 2002) outlines a protocol for obtaining voice samples and rating their voice quality. It was developed as a standard voice protocol based on expert consensus and psychophysically appropriate measurement of auditory perceptual qualities of voice. The CAPE-V has since obtained widespread research and clinical use, but research suggests considerable variability in how both expert and new clinicians use its rating scales. In this paper, I review remaining challenges to standardizing voice quality evaluation and describe ongoing research addressing these challenges.

Introduction

Since its development in 2002, the Consensus Auditory Perceptual Evaluation of Voice (CAPE-V; American Speech-Language-Hearing Association, 2002) has become a widely used protocol and measurement tool for evaluating voice. Its purpose was threefold: (1) to standardize terminology used by voice clinicians and scientists to describe voice quality; (2) to serve as a standard protocol for voice evaluations; and (3) to provide a psychometrically appropriate measurement tool for clinical evaluation of voice quality.

The CAPE-V is not a norm-referenced tool, but standardization of procedures was meant to ensure consistency among all clinicians conducting voice evaluations. Prior to the introduction of the CAPE-V, no single protocol outlined specific elicitation tasks or provided rating scales appropriate for auditory-perceptual evaluation of voice quality. Although the GRBAS (Hirano, 1981) was in widespread use, it provides only a 4-point Likert scale and has no associated scoring form. Considerable effort went into optimizing the validity and reliability of the chosen CAPE-V rating scale, to maximize the probability that a single voice would be rated in a similar way by different clinicians (inter-rater reliability) or by the same clinician on different occasions (intra-rater reliability; Kempster, Gerratt, Verdolini, Barkmeier-Kraemer, & Hillman, 2009).

Initial studies of the CAPE-V indicated sufficiently high inter-rater reliability and greater sensitivity to differences among voice samples than the GRBAS (Karnell et al., 2007; Nemr et al., 2012; Zraick et al., 2011). Unfortunately, despite these attempts to create a standard protocol, it appears that clinicians and researchers frequently do not follow the prescribed procedures (Lodhavia & Kempster, 2015; Marks, Nagle, & Heaton, 2015).

This paper reviews the continuing challenges of auditory-perceptual measurement of voice quality, particularly as they pertain to the CAPE-V. It then describes several lines of research meant to address these challenges.

Challenges to Auditory-Perceptual Evaluation of Voice

Perception of voice quality is often what drives an individual to seek a voice evaluation. Consequently, auditory-perceptual judgment is a critical outcome measure of voice quality that complements acoustic and aerodynamic measures. Perceptual outcomes rely on the subjective experience of the rater, however, and have inherent challenges that can be difficult to surmount. These challenges can be categorized as signal-based, rater-based, and task-based. Factors most relevant to the CAPE-V are described below.

Signal-Based Effects

A major factor affecting voice quality ratings is the multidimensionality of voice itself. The six quality features to be rated on the CAPE-V were chosen as the most commonly researched and most meaningful to clinicians and researchers. Judging pitch and loudness is relatively straightforward, and ratings can easily be compared to objective acoustic measures. The parameters of breathiness, roughness, strain, and overall severity have considerable perceptual overlap, making it difficult to separate and rate them individually. A “strained” voice, for example, may contain a component of roughness or breathiness. Should a clinician attempt to perceptually isolate and rate the dimension of strain without consideration of other voice parameters? Or are roughness and breathiness perceptually inextricable from strain? Individual raters take different approaches to quantifying the multidimensionality of voice, blunting inter-rater reliability for all but ratings of overall severity.

Similarly, the concurrent validity of loudness and pitch is easy to establish. Perceptions may differ, but ratings can be directly compared to acoustic measures of fundamental frequency and intensity. In the absence of a physically measurable correlate for the varying degrees of roughness, breathiness, strain, and overall severity, however, perceptual ratings rely on comparison to either an external standard (e.g., anchor sample) or a rater’s internal standard for each parameter.

Rater-Based Effects

All raters have a unique internal standard of voice quality against which they judge any voice sample. These standards are also unstable, particularly if the rater is exposed to a wide range of speakers and speech or voice types. Over time, clinicians adapt to disordered voice in general and to specific speakers’ voices in particular (Kreiman, Gerratt, & Precoda, 1990). Within-rater adaptations can also occur with exposure to different samples over a short period of time. In other words, listening to a number of voices over the course of a day can reset a clinician’s internal standard, the basis for all of perceptual judgments. These rater-based differences can cause a clinician to unknowingly apply different criteria to the same voice sample over time, resulting in inconsistent or unreliable absolute ratings of voice quality.

Task-Based Effects

Task factors affecting perceptual measurement relate to the task of rating (e.g., instructions), characteristics of the scale (e.g., anchors) and type of scale itself (e.g., categorical, ordinal, interval). For example, unless task instructions are very clear, raters may use quite different criteria for the same scale, particularly for multidimensional parameters of voice quality (Shrivastav, Sapienza & Nandur, 2005). As mentioned above, some clinicians attempt to isolate specific perceptual parameters when rating them; others include the effects of one parameter on their ratings of another. In this way, although they may conceptually have the same understanding of strain, roughness or breathiness, they are implementing this knowledge differently when providing ratings of voice quality. Clinicians represent an essential component of a measurement tool like the CAPE-V, and they must have a shared understanding of the task to provide consistent ratings. For these reasons, in addition to shared terminology, it is essential that all raters share an understanding of the instructions and of the perceptual dimensions they are being asked to rate.

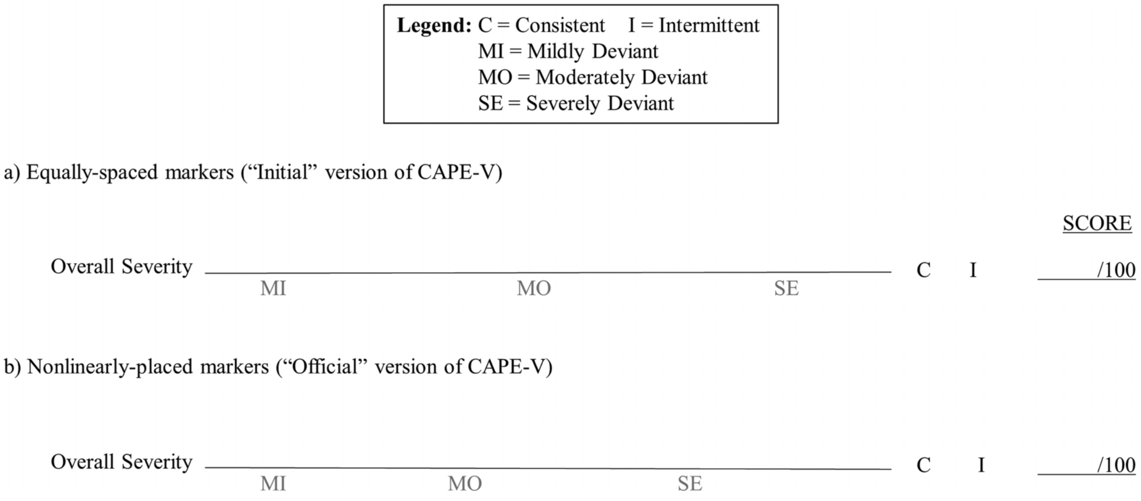
Characteristics of the rating task may therefore affect the consistency of raters. Rater consistency is also critical for interpreting perceptual ratings, and is reported in complementary ways: correlation (usually referred to as *reliability*) and agreement. Reliability refers to the extent to which a measurement is consistent and free of systematic or random error (Portney & Watkins, 2000). Intra- and inter-rater reliability reflect only the stability of the relative relationships among ratings; reliability coefficients are not affected by systematic errors. This means that there can be perfect correlation between data sets in which none of the points agree. This can occur if listener "drift" occurs between ratings sessions or if raters use a scale in parallel ways but at different levels (Kreiman, Gerratt, Kempster, Erman, & Berke, 1993). Rater *agreement* is a better way of reporting the extent to which obtained measures are actually the same. High levels of agreement suggest that raters used the same strategy or had the same internal standard for their judgments. It is conventional to accept ratings of plus or minus 1 for an EAI scale or within 1 centimeter on a VAS scale, but exact agreement may also be reported. Good agreement does not ensure good reliability, however; if the range of stimuli is small or if raters use only a limited area of the scale space, the consequent restricted range of ratings may reduce reliability. When it comes to applying CAPE-V findings across clinicians (or across research studies), strong inter-rater agreement is paramount. For example, agreement among raters is essential for creating a training paradigm with representative voice samples.

Using the appropriate type of rating scale is also critical for obtaining valid and reliable perceptual ratings that are sensitive enough to detect change. As an ordinal scale, the GRBAS does not accommodate degrees of difference among voices judged as "moderate;" there is no measurable difference between the four points of the scale. The GRBAS thus lacks the sensitivity to detect potentially meaningful perceptual differences among samples (Nemr et al., 2012).

Interval scales are used when the distance between two points can be quantified. Visual analog scales (VAS), in which these points are not marked, are frequently used to measure speech and voice quality (Bunton, Kent, Duffy, Rosenbek, & Kent, 2007; Yiu & Ng, 2004). The CAPE-V was created with a hybrid VAS; that is, the scales are traditional 100mm lines with labeled endpoints, and with additional points marked "mild," "moderate," and "severe" that are meant to serve as textual anchors. Adding textual labels to a VAS may dull the sensitivity of the scale, however, by giving it the appearance of a three-category scale. In other words, raters may subconsciously define categories for voice quality based on the position of the textual anchors (Nagle, Helou, Solomon, & Eadie, 2014). It is not clear that the statistical analyses used to analyze data obtained using a traditional VAS are appropriate for scales with textual anchors.

To complicate matters, an initial version of the CAPE-V was released with ordinal textual anchors placed at equal intervals ("equally spaced," see Figure 1a). There is substantial evidence from the psychophysical literature of systematic rater bias at the lower end of interval scales (e.g., Gescheider, 1997; Stevens, 1975), and the authors of the CAPE-V intended to account for this by placing textual markers nonlinearly (Kempster, 2007; Kempster et al., 2009). A version of the CAPE-V scales with nonlinearly placed textual markers does exist, with "severe" occupying a relatively larger portion of the scale than "mild" or "moderate" (Figure 1b). This scale accounts for systematic rater bias, but may yet have unexpected psychometric repercussions. Specifically, nonlinearly-placed textual labels suggest that intervals are known, but not equal. Ratings obtained using this type of scale are quite likely to systematically vary among raters in unknown ways, making interpretation difficult. A related issue should be noted: Both the "equally spaced" and "nonlinearly placed" versions of the CAPE-V are currently in use (and available on the internet). Care must be taken to use the same version if repeated ratings are obtained, to reduce the systematic error that may be introduced by using two scale types for the same protocol (Nagle et al., 2014).

Figure 1. Example Scales From Two Versions of the CAPE-V


Legend: C = Consistent I = Intermittent
MI = Mildly Deviant
MO = Moderately Deviant
SE = Severely Deviant
a) Equally-spaced markers ("Initial" version of CAPE-V)
SCORE
Overall Severity
MI
MO
SE
C
I
/100
b) Nonlinearly-placed markers ("Official" version of CAPE-V)
Overall Severity
MI
MO
SE
C
I
/100

In summary, although the CAPE-V advanced the cause of standardizing the terminology and tasks involved in evaluating voice, numerous issues remain to be addressed. First, raters appear to account for the multidimensionality of voice quality by using different rating strategies. Second, absolute ratings based on the unique internal standards of individual raters are not useful for comparison across raters, despite their high reliability; no "score" can be truly assigned to a voice. The CAPE-V, with no external standards, currently has no means of increasing agreement among raters. Third, it is not clear how best to analyze ratings obtained from the hybrid VAS used by the CAPE-V, particularly as different versions of the scale are in current use.

Addressing Challenges

In introducing the CAPE-V, its authors raised the possibility of creating exemplars for use as auditory anchors or for training purposes (Kempster et al., 2009). Funding was not allocated for this effort at the time of the CAPE-V's development, however, and auditory anchors have yet to be identified. Apart from funding such an effort, it is no small task to achieve consensus on univariate ratings of multidimensional voice parameters. Several lines of ongoing research address the challenges described above and promise to facilitate the development of auditory anchors and other features meant to further standardize the CAPE-V.

Signal-Based Effects

In recent years, researchers have synthesized voice samples exhibiting varying degrees of roughness and breathiness (Eddins, Vera-Rodriguez, Skowronski & Shrivastav, 2015; Patel, Shrivastav & Eddins, 2012a, b). Although strain has proven much harder to describe acoustically, research using the method of analysis by synthesis has provided insight about the variables that may affect perceived strain (Kreiman, Gerratt, Signorello & Rastifar, 2015). These efforts do not lessen the complexity of the voice signal, but they may provide a way of demonstrating concurrent validity for CAPE-V ratings. Samples synthesized to meet acoustic criteria could not only corroborate perceptual ratings, but also serve as auditory anchors in an eventual electronic version of the CAPE-V.

Rater-Based Effects

Two approaches to reducing rater-based variability may also strengthen the CAPE-V protocol. An ongoing investigation of how clinicians use auditory-perceptual scales takes a mixed methods approach and may provide insight into the basis of perceptual judgments. Asking clinicians how they interpret the terms and scales used in voice evaluation is a way to systematically collect authentic communicative activities in context. Combining qualitative data with contemporaneous ratings of dysphonia will provide clear links between research and clinical practice. It seems likely that these combined techniques will reveal two or more distinct ways in which clinicians approach auditory-perceptual ratings; for example, some may report trying to separate out breathiness from strain, whereas others may report including the effect of breathiness on any other parameter when rating a sample. It may then be possible to make these rater difference apparent on the CAPE-V form (e.g., “strain includes/does not include breathiness”), in the way that current versions include an option for “consistent/inconsistent” degree of a given voice parameter. Qualitative data may also provide a rationale for refined protocol instructions or modifications to the scale.

Given the variability with which the CAPE-V is used by clinicians, the need for training is clear. Training to the task of rating and to the sample type has been shown to improve agreement (Chan & Yiu, 2002; Eadie & Baylor, 2006). Ideally, a training paradigm would be appropriate for graduate students, but also for refreshing clinicians who use the CAPE-V only infrequently. The University of Wisconsin (UW) created a website meant to provide practice using the CAPE-V rating protocol to rate voice samples (https://csd.wisc.edu/slpgames/sims.html; Connor, Bless, Dardis, Vinney, & DoIT Engage Project Team, 2008). Case histories and voice samples from 45 speakers are provided, along with CAPE-V ratings obtained from a single expert clinician. This effort offers an excellent introduction to the CAPE-V and auditory-perceptual evaluation of voice, but an ideal training protocol would provide expert feedback from numerous raters. Maximizing inter-rater agreement of CAPE-V ratings in general would allow training protocols like the UW website to supply consensus-based feedback to learners.

Task-Based Effects

If consensus can be reached for training purposes, it should then be possible to create auditory anchors for use with the CAPE-V. Rating samples using CAPE-V is currently a paper and pencil task, but all of the scale types described above can be replicated in electronic form. It seems likely that the next version of the CAPE-V could be electronic, in which case auditory anchors could provide an external standard against which raters could judge any voice sample (Kreiman et al., 1993). Numerous questions remain to be addressed (e.g., How many anchors should there be? Should there be more at the more severe end of the scale? Does this differ for different voice parameters?). The benefits of providing an external standard to raters of any experience level justify this effort.

Considerable effort went into the initial creation of the hybrid VAS used in the CAPE-V, and it may well be the best way to obtain auditory-perceptual ratings of voice. However, the effects of textual anchors on clinicians using a hybrid VAS are unknown. If ratings of the same samples differ when obtained on a traditional VAS and a hybrid, then we need to investigate whether or not our current methods of analyzing these data are appropriate. The assumption that the distance between assumed points on a VAS is perceptually equal (i.e., 1mm is the same at any point on the scale) may be incorrect, particularly if a textual anchor is present.

Conclusion

The CAPE-V was created to provide a standard protocol, terminology, and rating scale for auditory-perceptual evaluation of voice. Despite considerable efforts to maximize the reliability and validity of ratings obtained using the CAPE-V, there is evidence of wide variability in clinician use of the protocol. Ongoing research is investigating several factors that have hindered development of auditory anchors and a CAPE-V training protocol.

References

American Speech-Language-Hearing Association. (2002). *Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V).* ASHA Special Interest Division 3, Voice and Voice Disorders. Retrieved from http://www.asha.org/uploadedFiles/ASHA/SIG/03/CAPE-V-Procedures.pdf

Bunton, K., Kent, R. D., Duffy, J. R., Rosenbek, J. C., & Kent, J. F. (2007). Listener agreement for auditory-perceptual ratings of dysarthria. *Journal of Speech, Language & Hearing Research, 50,* 1481–1495.

Chan, K. M., & Yiu, E. M. (2002). The effect of anchors and training on the reliability of perceptual voice evaluation. *Journal of Speech, Language, and Hearing Research, 45,* 111–126. doi: 10.1044/1092-4388 (2002/009)

Connor, N. P., Bless, D., Dardis, C., Vinney, L., & DoIT Engage Project Team. (2008). Simulations: Consensus Auditory-Perceptual Evaluation of Voice [CAPE-V] [Game/Simulation]. Retrieved from https://csd.wisc.edu/slpgames/sims.html

Eadie, T. L., & Baylor, C. R. (2006). The effect of perceptual training on inexperienced listeners' judgments of dysphonic voice. *Journal of Voice, 20,* 527–544.

Eddins, D. A., Vera-Rodriguez, A., Skowronski, M. D., & Shrivastav, R. (2015). Behavioral and computational estimates of breathiness and roughness over a wide range of dysphonic severity. *The Journal of the Acoustical Society of America, 138*(3), 1809. doi: 10.1121/1.4933740

Gescheider, G. (1997). *Psychophysics: The fundamentals* (3rd ed.). Mahwah N.J.: L. Erlbaum Associates.

Hirano, M. (1981). "GRBAS" scale for evaluating the hoarse voice & frequency range of phonation. In M. Hirano (Ed.), *Clinical examination of voice* (Vol. 5, pp. 83–84, 88–89). New York, NY: Springer-Verlag/Wien.

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.

Kempster, G. B. (2007). CAPE-V: Development and future direction. *Perspectives on Voice and Voice Disorders, 17,* 11–13. doi: 10.1044/vvd17.2.11

Kempster, G. B., Gerratt, B. R., Verdolini Abbott, K., Barkmeier-Kraemer, J., & Hillman, R. E. (2009). Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol. *American Journal of Speech-Language Pathology, 18,* 124–132.

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 & Hearing Research, 36,* 21–40.

Kreiman, J., Gerratt, B. R., & Precoda, K. (1990). Listener experience and perception of voice quality. *Journal of Speech & Hearing Research, 33,* 103–115.

Kreiman, J., Gerratt, B. R., Signorello, R., & Rastifar, S. (2015). Sufficiency of a four-parameter spectral model of the voice source. *The Journal of the Acoustical Society of America, 137,* 2266. doi: 10.1121/1.4920269

Lodhavia, A., & Kempster, G. B. (2015, November). *Fidelity to the CAPE-V.* Poster presented at the annual meeting of the American Speech-Language-Hearing Assocation, Denver, CO.

Marks, K. L., Nagle, K. F., & Heaton, J. T. (2015, June). *Improving agreement between graduate students and experts for CAPE-V measures.* Poster presented at the Voice Foundation Annual Symposium, Philadelphia, PA.

Nagle, K. F., Helou, L. B., Solomon, N. P., & Eadie, T. L. (2014). Does the presence or location of graphic markers affect untrained listeners' ratings of severity of dysphonia? *Journal of Voice, 28,* 469–475. doi: 10.1016/j.jvoice.2013.12.011

Nemr, K., Simões-Zenari, M., Cordeiro, G. F., Tsuji, D., Ogawa, A. I., Ubrig, M. T., & Menezes, M. H. M. (2012). Grbas and cape-v scales: High reliability and consensus when applied at different times. *Journal of Voice, 26,* 812.e17–812.e22. doi: 10.1016/j.jvoice.2012.03.005

Patel, S., Shrivastav, R., & Eddins, D. A. (2012a). Developing a single comparison stimulus for matching breathy voice quality. *Journal of Speech, Language & Hearing Research, 55,* 639–647. doi: 10.1044/1092-4388(2011/10-0337)

Patel, S., Shrivastav, R., & Eddins, D. A. (2012b). Identifying a comparison for matching rough voice quality. *Journal of Speech, Language, and Hearing Research, 55,* 1407–1422. doi: 10.1044/1092-4388(2012/11-0160)

Portney, L., & Watkins, M. (2000). *Foundations of clinical research: Applications to practice* (2nd ed.). Upper Saddle River: Prentice Hall Health.

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 & Hearing Research, 48,* 323–335.

Stevens, S. (1975). *Psychophysics: Introduction to its perceptual, neural, and social prospects*. New York, NY: Wiley.

Yiu, E., & Ng, C. (2004). Equal appearing interval and visual analogue scaling of perceptual roughness and breathiness. *Clinical Linguistics & Phonetics, 18,* 211–229.

Zraick, R. I., Kempster, G. B., Connor, N. P., Thibault, S., Klaben, B. K., Bursac, Z., . . . Glaze, L. E. (2011). Establishing validity of the consensus auditory-perceptual evaluation of voice (CAPE-V). *American Journal of Speech-Language Pathology, 20,* 14–22.

History:
Received December 01, 2015
Revised January 29, 2016
Accepted February 24, 2016
doi:10.1044/persp1.SIG3.47