Perceptual Assessment of Voice Quality: Past, Present, and Future

Article *in* Perspectives on Voice and Voice Disorders · July 2010

DOI: 10.1044/vvd20.2.62

2 authors:

Jody Kreiman
University of California, Los Angeles
187 PUBLICATIONS **3,662** CITATIONS
SEE PROFILE

Bruce Gerratt
University of California, Los Angeles
157 PUBLICATIONS **3,647** CITATIONS
SEE PROFILE

Perceptual Assessment of Voice Quality: Past, Present, and Future

Jody Kreiman and Bruce R. Gerratt

Department of Head and Neck Surgery, University of California School of Medicine
Los Angeles, CA

Abstract

Despite many years of research, we still do not know how to measure vocal quality. This paper reviews the history of quality assessment, describes some reasons why current approaches are unlikely to be fruitful, and proposes an alternative approach that addresses the primary difficulties with existing protocols.

Introduction

Measurement of voice quality is at the heart of clinical assessment of voice disorders. Vocal quality is a central concern to patients, who typically do not consider themselves improved until their voice sounds better. In fact, quality measures are arguably better for documenting treatment progress and assessing treatment efficacy than other kinds of measures, because they directly address the issue that led the patient to seek treatment in the first place.

The study of vocal quality (not surprisingly) has a long history, dating back at least to the Romans. In the oldest and most common approach to quality measurement, a listener rates the voice of interest on a set of scales that measure the listener's auditory impressions. The scales usually used to describe and assess quality are ingrained in Western culture and have changed very little in 2000 years. Familiar terms like *harsh, clear, bright, smooth, weak, shrill, deep, dull, thin, hoarse,* and *metallic* can be found in Roman writings on oratory (Austin, 1806) and also in modern studies of voice quality (e.g., Gelfer, 1988; Karnell et al., 2007). Centuries of consistent use give such scales a ring of truth that reinforces the widespread belief in their validity.

A number of rather similar rating scale protocols for quality evaluation have been proposed over the years. For example, users of the CAPE-V (Consensus Auditory-Perceptual Evaluation—Voice) protocol (Kempster, Gerratt, Verdolini Abbott, Barkmeier-Kraemer, & Hillman, 2009) rate voices on visual analog scales for overall severity, roughness, breathiness, strain, pitch, loudness, and any additional scales the clinician may wish to add. The GRBAS protocol (Hirano, 1981) assesses voices on scales for essentially the same qualities [Grade (equivalent to overall severity), Roughness, Breathiness, Asthenicity (weakness), and Strain, but uses four-point scales instead of marks on a line. Such protocols bear a surprising resemblance to the scales described in the 2nd century AD by Julius Pollux (Austin, 1806; see also Kreiman, Vanlancker Sidtis, & Gerratt, 2005). This approach is familiar, easy to apply, and easy to understand; further, resemblances among protocols and the lack of change over decades have sometimes created the impression that the major issues surrounding quality measurement have been solved and that clinicians and researchers have agreed on the best

way to measure quality in the clinic. Labeling the CAPE-V protocol a "consensus model" enhances this impression.

However, this apparent agreement also can be viewed as a failure to move beyond the longstanding familiar descriptive framework provided by the ancients. By conducting research using terminology and concepts derived from tradition and not from a theory or model of voice quality, users are left with no basis for answering some rather fundamental questions governing selection and interpretation of scales. For example, which scales should be included (or excluded) and why? How are scales related to underlying acoustic variables, and what is the relationship of one scale to the others (correlated? independent? interacting?)? Behind these issues lurks the larger problem of the ontological status of a protocol for assessing voice quality. Rating scale protocols for quality assessment never have been based on a model of voice quality, and, because the construct being measured is not well-defined, it is not possible to determine that a given set of scales or acoustic measurements is the "correct" one for measuring it (or even to address this question experimentally). This fundamental problem of validity—what we are really measuring and why—continuously has plagued attempts to make clinically meaningful measures of voice quality; but, without a clear delimitation of the thing being measured, this problem never will be solved, because it is impossible to determine that any proposed scale or set of scales is either necessary or sufficient.

One solution to this problem may lie in the American National Standards Institute (ANSI) definition of "quality" as those attributes of a sound other than its pitch and loudness that allow a listener to judge that two sounds are the same or different (ANSI, 1960). This definition implies that the goal of a quality assessment protocol is to quantify not a set of specific vocal attributes, but instead the complete sound of a voice—the speaker's integral, overall, personal vocal quality. By applying this definition, it becomes possible to evaluate the adequacy of a quality assessment tool: To the extent that a set of ratings or acoustic measures combine with measures of pitch and loudness to precisely specify the overall, integral quality of the target stimuli (and not just a subset of specific features), that assessment protocol is a valid measure of quality. Existing rating scale protocols cannot meet this standard. A set of GRBAS or CAPE-V ratings (or any other set of ratings) gives users no idea whatsoever about the integral quality of the voice being measured (nor were they intended to). Because the thing being measured is unspecified in such protocols, it is difficult to imagine how this limitation can be overcome.

An additional longstanding problem is the issue of rating reliability in quality assessment protocols. Scale reliability in the traditional statistical sense has very little meaning in clinical settings, because it measures the likelihood that a new random sample of raters would produce the same mean rating as the group studied, averaged across all the voices studied. This is seldom what we want to know in the clinic or the lab, where the interesting question is the likelihood that two individual raters will produce the same rating for a given voice sample: the probability of listener agreement. These two approaches are quite different and can lead to very different conclusions about reliability. Even when intraclass correlations or Cronbach's alpha values suggest data sets are completely reliable, the likelihood of listener agreement still can be very poor, sometimes failing to reach even chance levels in the mid-range of the rating scale (Kreiman & Gerratt, 1998).

Clinicians and researchers have suggested a number of solutions to the "unreliable rater" problem. One of these (Shrivastav, Sapienza, & Nandur, 2005) involves averaging ratings to achieve a reliable mean, based on the assumption that rating variability across and within raters is mostly random. Other researchers have suggested training listeners to increase the extent to which they share common standards for different qualities (e.g., Chan & Yiu, 2006; Shewell, 1998); using fewer scale values (which increases agreement, but also increases the likelihood of chance agreement; e.g., De Bodt, Wuyts, Van de Heyning, & Croux, 1997; Webb et al., 2004); and applying anchored protocols with comparison stimuli to which the target voices

are compared (thereby—hypothetically—reducing reliance on internal standards for different qualities; e.g., Awan & Lawson, 2008; Chan & Yiu, 2002; Gerratt, Kreiman, Antoñanzas-Barroso, & Berke, 1993). Instead of attempting a detailed or quantitative assessment of voice quality, other authors have suggested that primary outcome measures should focus on asking patients about satisfaction with their voices or voice-related quality of life (e.g., Hogikyan & Sethuraman, 1999; Hogikyan, Wodchis, Terrell, Bradford, & Esclamado, 2000; Franco & Andrus, 2009). This last approach helps clinicians to gauge the overall success of a treatment approach; however, satisfaction ratings do not provide insight into why the treatment succeeds or fails with individual patients or about which physical changes actually modified the resultant sound.

Another popular solution to the problem of rater (un)reliability is to substitute instrumental measures for subjective assessments of voice quality. Knowledge of what physical and acoustic changes actually cause which variations in quality (and vice versa) would provide a basis for analysis protocols that help clinicians understand the reasons for success and failure in clinical cases, greatly enhancing their ability to focus and evaluate treatment. This standard is not met by current protocols, such as the Dysphonia Severity Index (Wuyts et al., 2000), the Hoarseness Diagram [Frohlich, Michaelis, Strube, & Kruse, 2000], the Multidimensional Voice Program (MDVP; Kay Elemetrics, 1993), and the Acoustic Voice Quality Index (Maryn, Corthals, van Cauwenberge, Roy, & de Bodt, in press). The reason is, again, they are not linked to any model of vocal quality, so that changes in a measure are not understood easily in terms of changes in the sound of the voice. The perceptual salience of the parameters measured by these systems is often unknown or known to be very limited. For example, listeners apparently are quite insensitive to changes in jitter and shimmer in sustained vowels (Kreiman & Gerratt, 2005); these measures nonetheless are included in almost every current acoustic voice assessment protocol. Virtually no psychometric work has been undertaken to establish the relationship between voice quality and acoustic measures like (to name a few) cepstral peak prominence (Hillenbrand, Cleveland, & Erikson, 1994), amplitude differences between the first harmonic and the first or second formant (H1-A1 or H1-A2; Hanson, 1997), or the soft phonation index (Roussel & Lobdell, 2006), other than reports of varying correlations between an acoustic measure and some rated voice quality. Because we cannot evaluate the usefulness of an instrumental measure of quality without assessing quality as part of that process, objective voice analyses do not solve the problem of rater unreliability or rating scale insufficiency. Rather, they introduce new questions about their own validity, because their status as measurements of quality perception derives solely from correlations, rather than evidence derived from psychoacoustic experimentation.

In contrast to these approaches, we envision modeling quality psychoacoustically as a set of acoustic parameters that combine non-redundantly to capture the complete, integral voice pattern—the way a person sounds. To derive such a model, we must first develop a reliable and valid method of measuring listeners' perceptions as we manipulate acoustic parameters to establish valid links between vocal acoustics and perceived quality. We have made progress on this front, starting with studies seeking the sources of variability in listeners' judgments of quality in rating protocols, to guide development of alternative measurement tasks that are specifically designed to facilitate agreement among listeners. Four factors have been identified that account for most of the variability in perceptual measurements of quality: difficulties isolating individual attributes in complex acoustic voice patterns (the most important factor); instability of listeners' internal standards for different qualities; measurement scale resolution; and the magnitude of the attribute being measured (Kreiman, Gerratt, & Ito, 2007). When these factors were controlled through use of a method-of-adjustment task, listeners agreed exactly 96% of the time in their quality assessments.

Once reliable quality assessments can be obtained, we next can determine which acoustic parameters evoke perceptible changes, thereby identifying a set of acoustic measures

of voice quality that are perceptually valid (in the same way that the decibel is a perceptually valid acoustic measure of sound intensity by virtue of its well-defined relationship to perceived loudness). Using this process, we can ultimately derive a truly psychoacoustic model of voice quality that specifies a set of acoustic measures that combine to quantify the sound of an individual voice (and not just the magnitude of certain attributes). Such a model could eliminate the need for subjective quality measures (including verbal scale ratings), because the perceptual importance of each acoustic parameter can be established; interactions among parameters can be modeled; and the composite set of parameters can be selected to specify voice quality. In other words, in the context of a model of quality, the parameters are not arbitrary; they are not redundant, and they are all perceptually valid.

How might one go about deriving such a measurement instrument? If we assume that the amount of attention listeners pay to an attribute is roughly proportional to the extent to which that attribute varies across voices, we can begin by first identifying acoustic attributes that best differentiate voices and then test the perceptual significance of these parameters. When we have identified a set of parameters for which listeners can hear small differences in a variety of contexts, we can apply speech synthesis to determine whether those parameters, plus pitch and loudness, are adequate to capture the quality of an individual's voice. Through iterations of this process, we can identify parameters that are redundant or that interact with each other perceptually. For example, the perceptual importance of spectral noise depends on the amount of harmonic energy present in the higher part of the spectrum: the more harmonic energy is present, the more change in inharmonic energy is needed before listeners can hear a difference (Gerratt & Kreiman, 2001; Shrivastav et al., 2005). Thus, a perceptually valid scale for spectral noise levels will need to correct for this interaction, in the same way that measures of loudness correct for perceptual interactions with frequency via "equal loudness" curves (Fletcher, 1934).

The ultimate goal of these psychoacoustic modeling efforts is identification of a small set of acoustic parameters that are necessary and sufficient to precisely reproduce the overall, integral quality of voices and are, therefore, suitable to evaluate changes in voice quality in the clinic and elsewhere. Such measures would constitute a valid assessment tool, because they are objective measures whose relationship to quality is understood theoretically, and whose relevance to quality is demonstrated unequivocally by psychoacoustic evidence. This tool would allow clinicians to interpret acoustic changes in terms of their associated perceptual analog, in the same way that the sone relates loudness and sound intensity (Stevens, 1936) and the mel relates acoustic frequency and perceived pitch (Stevens, Volkmann, & Newman, 1937). Once this goal is achieved, we can consider the even more ambitious goal of a comprehensive theory of voice that relates changes in vocal physiology to the resultant changes in voice quality and, conversely, maps changes in quality to the physical changes that caused them. Such a model would provide a theoretical basis for clinical assessment, because it would specify causal links from laryngeal physiology, to voice acoustics, to quality, and back. We submit that development of such a comprehensive theory should be the primary goal of voice research.

In conclusion, the last 2,000 years have produced awareness and descriptions of the importance of voice and its uses, but previous work has not led to very much theoretical understanding of the "whys" of voice quality. As a result, no foundation of psychoacoustic evidence exists for the development or validation of measurement techniques. Voice quality never has been studied as part of a broader theory that encompasses the whole speech chain, so we cannot predict changes in the sound of a voice with changes in vocal fold vibrations or interpret a change in the sound of a voice when one occurs. However, we may be nearing a solution to the long-term problem of generating reliable and valid measures of voice, whose derivation will make it possible to address other critical, basic questions that await our attention in the future.

References

American National Standards Institute. (1960). *Acoustical terminology. ANSI S1.1.12.9.* New York, NY: Author.

Austin, G. (1806). *Chironomia.* London: Cadell and Davies. Reprinted by Southern Illinois University Press, Carbondale, IL, 1966.

Awan, S. N., & Lawson, L. L. (2008). The effect of anchor modality on the reliability of vocal severity ratings. *Journal of Voice, 23,* 341-352.

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

Chan, K. M. K., & Yiu, E. M.-L. (2006). A comparison of two perceptual voice evaluation training programs for naive listeners. *Journal of Voice, 20,* 229-241.

De Bodt, 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,* 74-80.

Fletcher, H. (1934). Loudness, pitch, and the timbre of musical tones and their relation to the intensity, the frequency, and the overtone structure. *Journal of the Acoustical Society of America, 6*, 59–69.

Franco, R. A., & Andrus, J. G. (2009). Aerodynamic and acoustic characteristics of voice before and after adduction arytenopexy and medialization laryngoplasty with GORE-TEX in patients with unilateral vocal fold immobility. *Journal of Voice, 23*, 261-267.

Frohlich, M., Michaelis, D., Strube, H. W., & Kruse, E. (2000). Acoustic voice analysis by means of the hoarseness diagram. *Journal of Speech, Language, and Hearing Research, 43*, 706-720.

Gelfer, M. P. (1988). Perceptual attributes of voice: Development and use of rating scales. *Journal of Voice, 2*, 320-326.

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

Gerratt, B. R., & Kreiman, J. (2001). Measuring vocal quality with speech synthesis. *Journal of the Acoustical Society of America, 110*(5), 2560-2566.

Hanson, H. M. (1997). Glottal characteristics of female speakers: Acoustic correlates. *Journal of the Acoustical Society of America, 101,* 466-481.

Hillenbrand, J., Cleveland, R. A., & Erickson, R. L. (1994). Acoustic correlates of breathy vocal quality. *Journal of Speech and Hearing Research, 37*, 769-778.

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

Hogikyan, N. D., & Sethuraman, G. (1999). Validation of an instrument to measure voice-related quality of life (V-RQOL). *Journal of Voice, 13*, 557-569.

Hogikyan, N. D., Wodchis, W. P., Terrell, J. E., Bradford, C. R., & Esclamado, R. M. (2000). Voice-related quality of life (V-RQOL) following type I thyroplasty for unilateral vocal fold paralysis. *Journal of Voice, 14,* 378-386.

Karnell, M. P., Melton, S. D., Childes, J. M., Coleman, T. C., Dailey, S. A., & Hoffman, H. T. (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.

Kay Elemetrics. (1993). *Multi-Dimensional Voice Program* (MDVP). [Computer program.] Pine Brook, NJ: Author.

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. (1998). Validity of rating scale measures of voice quality. *Journal of the Acoustical Society of America, 104*, 1598-1608.

Kreiman, J., & Gerratt, B. R. (2005). Perception of aperiodicity in pathological voice. *Journal of the Acoustical Society of America, 117*, 2201-2211.

Kreiman, J., Gerratt, B. R., & Ito, M. (2007). When and why listeners disagree in voice quality assessment tasks. *Journal of the Acoustic Society of America, 122*, 2354-2364.

Kreiman, J., Vanlancker-Sidtis, D., & Gerratt, B. R. (2005). Perception of voice quality. In D. B. Pisoni & R. E. Remez (Eds.), *Handbook of speech perception* (pp. 338-362). Walden, MA: Blackwell.

Maryn, Y., Corthals, P., van Cauwenberge, P., Roy, N., & de Bodt, M. (in press). Toward improved ecological validity in the acoustic measurement of overall voice quality: Combining continuous speech and sustained vowels. *Journal of Voice.*

Roussel, N. C., & Lobdell, M. (2006). The clinical utility of the soft phonation index. *Clinical Linguistics and Phonetics, 20,* 181-186.

Shewell, C. (1998). The effect of perceptual training on ability to use the Vocal Profile Analysis scheme. *International Journal of Language & Communication Disorders, 33*, 322-326.

Shrivastav, R., Sapienza, C., & 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.

Stevens, S. S. (1936). A scale for the measurement of a psychological magnitude. *Psychological Review, 33*, 405-416.

Stevens, S. S., Volkmann, J., & Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America, 8*, 155-210.

Webb, A. L., Carding, P. N., Deary, I. J., MacKenzie, K., Steen, N., & Wilson, J. A. (2004). The reliability of three perceptual evaluation scales for dysphonia. *European Archives of Oto-Rhino-Laryngology, 261*,429-434.

Wuyts, F. L., De Bodt, M. S., Molenberghs, G., Remacle, M., Heylen, L., Millet, B., et al. (2000). The dysphonia severity index: An objective measure of vocal quality based on a mutiparameter approach. *Journal of Speech, Language, and Hearing Research, 43,* 796-809.

Author Note

A preliminary version of this paper was presented at the 2009 ASHA annual Convention. The research described was supported by grant DC01797 from the National Institutes on Deafness and Other Communication Disorders.