PHONOSCOPE™

THE PERCEIVED ROLE OF VOICE PERCEPTION IN CLINICAL PRACTICE

Moderator
Robert F. Orlikoff, Ph.D.

PHONOSCOPE Content Editor, Instrumentation

Participants
James Dembowski, Ph.D.
James Fitch, Ph.D.
Marylou Pausewang Gelfer, Ph.D.
Bruce R. Gerratt, Ph.D.
John A. Haskell, Ed.D.
Jody Kreiman, Ph.D.
Dale Evan Metz, Ph.D.
Nicholas Schiavetti, Ph.D.
Ben C. Watson, Ph.D.
Virginia Wolfe, Ph.D.

Several expert clinicians and researchers were asked to address whether it is appropriate to use auditory-perceptual judgments of voice quality in the assessment and management of dysphonia, and if so, under what circumstances. Participants commented on the relative advantages and disadvantages of procedures to "quantify" listener-perceived voice characteristics and on whether this has become a moot issue now that clinically feasible ways to objectively measure the "quality" of the acoustic voice product and to identify impairment in vocal physiology are available. Although the roundtable participants differed substantially in their opinions, the collection of arguments presented provide a thorough and thought-provoking overview of a topic that is likely to become increasingly important to all professionals involved in the rehabilitation of vocal disorder.

The trained clinical ear has been the instrument of choice for generations of voice specialists. This continues to be true for a large number of voice clinicians, whose clinical decision-making may be directed solely by judgments derived from auditory impressions made while listening to a patient speak

PHONOSCOPE
Volume 2, Number 2, pp. 87–106

or vocalize. In fact, even when objective physical measures are at hand, judgments of voice quality are often held as the standard against which the clinical value of such data are determined. Be this as it may, the error and bias associated with unaided auditory perception is well known and have been studied and described in some detail by a number of voice researchers, teachers, and clinicians.[1-9] As such, there are those who eschew auditory impressions of voice quality for most clinical purposes, believing that, by and large, they provide an inadequate and untenable basis for outlining and guiding the course of intervention.

Never before has the voice specialist had access to a greater variety of clinical instrumentation than is available today. Such instruments, typically electronic and, of late, under computer control, provide the clinician with the means of obtaining relatively quick and accurate information about phonatory function and the acoustic result, often in an objective quantitative manner. The ready accessibility of these instruments, coupled with recent advances in our understanding of voice production and the growing importance of voice rehabilitation in the medical setting, has provided the impetus for a more physiologic approach to the management of dysphonia than has historically been the case. It has also called the primary importance usually ceded to auditory-perceptual judgments of voice quality into serious question.

For this roundtable, several noted experts were asked to contribute their thoughts on the issue of what role, if any, perceptual impressions of voice quality should play in the assessment and treatment of the dysphonic patient. Taking note of our present and ever-developing capability to objectively measure the "quality" of the acoustic voice product and to characterize specific impairments in vocal physiology, participants were asked to comment on both the feasibility and clinical utility of perception-based procedures to rate, scale, or otherwise "quantify" voice quality. Although not designed to be a debate, it was hoped that the collection of arguments and opinions offered would represent a lively and thought-provoking overview of a topic that is sure to receive increasing attention in the voice literature and in clinical practice.

MARYLOU PAUSEWANG GELFER, PH.D.

Department of Communication Sciences and Disorders
University of Wisconsin—Milwaukee

In a world in which there are increasingly available ways to acoustically and physiologically quantify aspects of voice and vocal production, does there remain a clinical need to measure perceived voice quality? In my opinion, despite our ever-growing list of clinical hardware and software, they do not provide the voice clinician with the tools necessary to diagnose and manage voice patients in the absence of perceptual measures.

Let us begin with the notion that voice is a multidimensional phenomenon, comprised of a number of elements that contribute to overall voice quality and vocal effectiveness. These elements often interact in ways that are difficult to measure objectively but are nonetheless important in clinical diagnosis and management. For example, a male-to-female transsexual subject in one of our perceptual studies had a speaking fundamental frequency of 180 Hz (very much in the female range) and formant frequencies in isolated vowels that closely approximated female norms. The objective acoustic measures made it seem as if the vocal transition had been successful. Yet, when listeners were asked to rate whether a tape of this speaker's voice was produced by a male or a female, 90% said "male."

Relatedly, a patient in our Voice Clinic complained of a "low" and "raspy" voice. When a spontaneous speech sample was analyzed acoustically, it was discovered that the patient's mean speaking fundamental frequency was, in fact, within normal limits. The patient's jitter, shimmer, and signal-to-noise ratio in isolated vowels also approximated that of the normal voice. Respiratory measures such as vital capacity indicated no abnormalities. These instrumental results give very little direction for intervention. However, careful perceptual analysis of the transcribed speech sample revealed that the "low" and "raspy" words and phrases were produced almost exclusively at the ends of sentences, which, from a linguistic standpoint, were overly long and phrased inappropriately. Clearly, there was an interaction between linguistic and prosodic variables, perhaps respiratory patterns, vocal fundamental frequency, and quality. Such interactions are difficult to determine through instrumental measures, which focus on discrete and isolated voice components by their nature. In voice evaluation, training, and therapy, the whole is greater than the sum of the parts, and the whole is better judged perceptually.

The above examples illustrate some of the difficulties with instrumental analysis of voice. First, available objective measures may not be measuring the salient attributes listeners use in judging voice. Second, the patient's concern about his or her voice is almost always based on spontaneous or

conversational speech. While we are currently able to make acoustic measures of frequency and intensity in connected speech (see Gelfer and Young[10] for a proposed method for the latter), most objective measures of voice "quality" are based on isolated vowel samples. Measures of "quality" or, more accurately, phonatory stability, such as jitter and shimmer, are based on the assumption that the patient is trying to maintain a steady vocal frequency. Spectrally-related measures, such as harmonics-to-noise ratio, are based on the assumption that the patient is trying to maintain a stable formant structure. Unfortunately, neither of these assumptions is true in connected speech, where frequency, formant structure, and voicing are constantly fluctuating in a purposeful manner. A patient may be very capable of maintaining stable phonation, but may have difficulty when producing connected speech. Furthermore, even for sustained vowels, it is often difficult to measure phonatory stability reliably. Gelfer,[11] for example, showed that jitter, shimmer, and signal-to-noise ratio can vary as a function of both the frequency and intensity of phonation. A host of other factors, including the examination room noise level and the type of microphone[12] and tape recorder[13] used for measurement, can influence results substantially. Given the difficulty of making reliable measures in laboratory-controlled conditions, it is somewhat unrealistic to expect these types of analyses to be valid and reliable in clinical situations. Moreover, even when well-calibrated and functional equipment is at hand, normative data are not always available or applicable to the clinician's particular measurement hardware and software system. For example, at least six or seven different jitter algorithms are used by various commercially-available analysis systems, and the normative data published using one algorithm may not be generalizable to another.[14] Although instrumental voice analysis may be the way of the future, at present it is not a reality for most clinicians.

Alternatively, perceptual scaling and other perceptual measures are readily available to the clinician and, if done conscientiously, can provide a global picture of the patient's vocal performance. Perceptual measures can also furnish significant information on how various vocal attributes may be affected by linguistic factors, prosody, and other variables, and may take several forms. It is also important that clinicians not overlook the patient's own perception of voice quality and how well it meets his or her vocational/social needs. Such outcome information is not amenable to objective assessment. While I regard perceptual measures as critically important to clinical practice, I also believe that objective acoustic and physiological measures are useful in diagnosing and treating voice disorders. Ideally, I would recommend that the clinician's perceptual judgments should be complemented by objective measures when possible and, if there is a discrepancy, its source should be explored.

P. H. DEJONCKERE, M.D., PH.D.

The Institute of Phoniatrics, Utrecht University Utrecht, The Netherlands

I am of the opinion that it is useful to quantify one's perception of voice quality using a simple and basic system that is standardized among clinicians throughout the world.

How?

In my institute (as in many others), we use the GRBAS scale[15] in routine clinical practice. We have found the most important items of the scale to be G (grade), R (roughness), and B (breathiness). These have largely shown good inter- and intra-rater agreement.[16] They also correspond well to the German "RBH" system (Rauhigkeit = roughness; Behauchtheit = breathiness; Heiserkeit = hoarseness/grade).[17] These items are relatively easy to define and are suitable to describe most types of dysphonia, with the possible exception of "substitution" voices and spasmodic dysphonia. Nonetheless, specific perception protocols need to be developed and verified. For instance, as a weighting factor, we have added the "I" (instability) parameter to the GRB scale that relates to perceived variability/fluctuation in voice quality over time. We believe that this parameter helps explain some of the discrepancies between acoustic analysis and overall perceptual judgment.[18]

In my opinion, perceptual evaluation, albeit subjective, can become a valid method to be used in efficacy studies. Using a very simple, basic scheme, such evaluation can be conducted in a systematic, quantitative way. It has been my experience that rating protocols developed by phoneticians, while useful for research purposes, are often not suitable for daily use in the voice clinic. My preference is to evaluate each parameter with a visual analog scale. The alternative is a numerical scale, such as 0—1—2—3. This was recently discussed exhaustively in Rome within the Committee on Phoniatrics of the European Laryngological Society. It is the charge of this committee to prepare guidelines for standardized

minimal voice evaluation, especially for publication of therapeutic results or new (phonosurgical) methods. It is also hoped that this will substantially facilitate meta-analyses.

Why?

There presently exists no single or complex objective parameter or measure that can precisely and accurately account for voice quality. Therefore, perception remains, in my opinion, the "gold standard." If we are concerned with voice pathology, it seems obvious to me that what is most meaningful to the patient (as soon as cancer, etc., has been ruled out) is how other people perceive his voice. It is especially important that the attention of the listener is not deviated from the content of the message by a "disturbing" voice quality. Perceptual rating is also the "gold standard" for assessing the clinical relevance of acoustical analyses. Thus, it is recommended, at least in published reports, that authors refer to a standard basic perceptual evaluation when describing a group of voice patients, when assessing the efficacy of a voice treatment regimen, or when proposing a new method, system, parameter, or protocol for voice assessment.

Voice quality needs to be quantified for medico-legal purposes. That is, the assessment of voice quality should correspond to the patient's voice complaint in terms of percent of impairment, disability, and handicap. It is important to compare what we hear, what we see, and what the patient perceives with acoustic analyses and measures of vocal efficiency.[19-21] This helps the clinician to understand the problem and to make a good global diagnosis and proposal for therapy, especially when a trial course of therapy is used to evaluate the need for surgical intervention.

VIRGINIA WOLFE, PH.D.

Auburn University at Montgomery
Montgomery, Alabama

JAMES FITCH, PH.D.

Auburn University
Auburn, Alabama

In clinical practice, one of the most important reasons for pursuing the development of objective measures of voice is a practical one: Objective measures impart credibility to clinical judgments. In general, there are three groups for whom objective data can be particularly helpful in the evaluation and treatment of voice disorders.

First is the patient and the patient's family, who deserve to be given all information that can help them understand the nature and extent of the disorder. Verbal descriptions of vocal disorder and the possible effects of intervention are the primary means of communicating this to patients. Credibility, however, will be enhanced if the clinician can provide objective data that quantify the results of intervention. Care must be taken to explain the meaning of the objective data and the extent to which it can be used to make judgments.

The second group for whom use of objective data can be useful is the professional community, in particular those in the medical profession. Medical personnel are accustomed to evaluating disorder by means of visual imaging and objective testing procedures that, by and large, yield quantified or quantifiable results. Objective measures of voice lend credibility to the verbal descriptions provided by clinicians.

The third group is made up of third-party funding sources, such as insurance, HMOs, and federal programs. The economic necessity of providing services must be established by the provider. Providers who use objective data to support their claims will be more credible.

Clinical judgments that should be supported by objective data include the categorization of different pathological voice qualities as well as the assessment of dysphonic severity. The combined results of studies that have simultaneously explored acoustic correlates of pathological qualities present some interesting similarities and differences. Several studies, for instance, have reported correlations between acoustic measures and perceived voice types, particularly for the qualities of "breathiness" and "roughness." In one such study, Eskenazi, Childers and Hicks[22] reported that the HNR (harmonics-to-noise ratio) and SFR (spectral flatness of the residue signal) were associated with "roughness," while percent jitter was associated with "breathiness." Martin, Fitch and Wolfe[23] and de Krom,[24] however, found the same acoustic measure, HNR, to be associated with both "breathiness" and "roughness." Takahashi and Koike[25] also found vocal shimmer to be associated with both perceived qualities, while, Wolfe, Fitch and Martin[26] determined that a combination of three time-domain measures (F_0 variation, peak F_0 variation, and F_0 tremor frequency) or NHR combined with any of five time-domain frequency perturbation measures

predicted judged "roughness," whereas only shimmer predicted perceived "breathiness." Therefore, a consideration of only this small group of studies—united by their common approach to the rating of two voice qualities—suggests the following:

1. it is not possible to categorize different dysphonic voice qualities based on the time and frequency-domain measures used in most studies; and
2. the severity of different dysphonic voice qualities can be predicted by the same acoustic measures.

With regard to the second point, most studies show only moderate to moderately high correlations suggesting that perceptual impressions include aspects that often may not be captured by acoustic measures. Therefore, the objective data that are currently available must be used cautiously. The ear of the clinician is by far the most valuable tool in the evaluation and treatment of voice disorders. Nevertheless, as discussed previously, there is ample reason for the profession to continue to identify meaningful objective measures that can be validated and standardized.

BRUCE R. GERRATT, PH.D.
JODY KREIMAN, PH.D.

Division of Head and Neck Surgery
UCLA School of Medicine
Los Angeles, California

The question at hand—Is there a clinical need to "measure" listener-perceived voice quality?—implies that our present ability to assess voice quality objectively may be sufficient, given currently used objective methods of acoustic and physiologic measurement. We would argue that:

1. documenting vocal quality is an essential element in the assessment, treatment, and management of a dysphonic patient;
2. vocal quality is an interaction between a signal and a listener, and therefore that its measurement can never reasonably be divorced from the listener; and
3. existing evidence does not support the equivalence of perceptual and objective measures of pathologic vocal quality, and thus does not justify the substitution of one class of measurement for another.

The need to measure any quantity or quality is dictated by the purpose at hand. In evaluating dysphonic patients, a clinician usually has two major goals. One is to discover the anatomic and physiologic sources of the laryngeal problem, which may lead directly to interventions to correct the disorder. The second is to measure the effects of the deviant laryngeal condition on the voice, including its quality, pitch, and loudness. These aspects of the voice disorder are usually the focus of patients' concerns. Patients' judgments of quality, pitch, and loudness initially lead them to seek care for a voice disorder, and they judge these variables during and after treatment in their assessment of treatment success. Except in the case of life-threatening diseases, judgments of these perceptual aspects of voice also contribute greatly to a clinician's decision to initiate and continue treatment. Thus, because the patient and the clinician place such great importance on perceptual aspects of voice, measurement of perceived quality must be viewed as a prime concern in patient management.

Despite the obvious clinical importance of voice quality, the appropriate method for measuring what listeners hear remains an unresolved issue. Many clinicians and researchers have recognized that listeners do not agree with each other about the extent to which voices possess perceived qualities such as "breathiness" and "roughness." We have written extensively on this subject, repeatedly finding significant differences among listeners in their perception of pathological voices. For example, we examined the likelihood that two listeners would agree in their ratings of a single voice on traditional scales like "breathiness" and "roughness".[27] Across scales, listeners agreed so poorly in their ratings of most voices that mean ratings could not be used fairly to represent the extent to which a voice possessed a quality. In fact, we did not find a single voice in all our data that listeners agreed was moderately deviant in quality (although listeners did sometimes agree that voices are normal or severely "breathy" or "rough"). Because mean ratings in the midrange of a scale apparently serve only to indicate that listeners disagree about the quality of a voice, traditional unidimensional scales for voice quality may not be adequate measures of what listeners hear when they listen to pathological voices.

Further evidence against the usefulness of traditional voice rating paradigms comes from multidimensional scaling studies. For example, results of one recent study[28] suggested that listeners do not share a common perceptual space for pathological voices. Familiar scales like "breathiness" and "roughness" failed to emerge from these studies. Instead, voices clustered together in groups that

lacked subjective unifying percepts, with different clusters emerging from each listener's data so that across listeners no two voices were consistently close together in the perceptual spaces. Thus the perceptual assessment of voice quality appears to be plagued by problems of scale validity, along with the issues of scale reliability described above.

Data like these leave clinicians faced with a dilemma. Perceptual assessment of voice quality is an essential component of the diagnosis and treatment of voice disorders, but apparently no acceptable protocol exists for this assessment. Much modern research on voice represents efforts to circumvent this problem by substituting "objective" measures of acoustics, physiological function, or airflow for ill-defined, unstandardized, and unreliable perceptual measures of dubious validity. In particular, acoustic measures continue to receive extensive attention from researchers, and are becoming the preferred means of documenting vocal quality in the clinic. These measures have long been popular in research applications, and the availability of off-the-shelf, automated programs now permits researchers and clinicians to generate acoustic values in almost real time. Acoustic measures have an additional attractive feature, in that they appear to provide a connection between the laryngeal physiology and the resulting perceived vocal quality, as in the classic speech chain model.[29]

However, a strong empirical link between measures must be established before one can be substituted for another. In the present case, acoustic measures are not particularly good indices of vocal physiology. The predictive relationship between physiology and vocal acoustics is rather weak, probably because of the complexity of these two links in the speech chain. For example, acoustic perturbation can arise from a great number of sources, including irregularities in muscular innervation, secretions or mass lesions on the vocal folds, tension asymmetries, randomness in the flow through the glottis, irregularities in source-vocal tract interactions through unstable articulatory configurations, or almost any other deviation of laryngeal function.[30] Because measured jitter, shimmer, or harmonics-to-noise ratio can be accounted for by many physiological conditions, acoustic perturbation measures provide very poor descriptions of the laryngeal behavior responsible for the perturbation.

In addition to their weak association with vocal physiology, acoustic measures are also poor indices of perceived vocal quality. Again considering perturbation, studies examining the correlation between perceptual qualities and jitter, shimmer, or noise have had consistently negative results[22,31,32] (see Ludlow et al.[33] for a review). Several studies have examined univariate relationships between acoustic measures and perceived "breathiness."[8,22,23,31,34-39] However, correlations are generally low to moderate, and no consistent relationship emerges between rated quality and any acoustic measure or family of measures. In particular, there are two kinds of variability that are apparent. First, many different acoustic measures may be correlated with one quality. For example, "breathiness" has been correlated with spectral slope, spectral noise, jitter, and shimmer. Secondly, a single acoustic measure may correlate strongly, moderately, and weakly with a single quality. For instance, in one study,[35] correlations between spectral noise and "breathiness" reportedly ranged from .14 to .72. Similar findings exist for other acoustic and perceptual measures (e.g., jitter and "roughness"; see Kreiman and Gerratt[40] for a review). Thus, the links between voice quality and measures of vocal acoustics are too weak and variable to justify substituting one kind of measure for the other.

Abandoning acoustic measures or combining them with aerodynamic, stroboscopic, electromyographic, or other direct or indirect indices of vocal function to represent perceived quality is not a viable solution to this problem at present, because the relationship between perceived quality and physiologically-based or aerodynamic measures is also not well understood. As in the case of acoustics, existing studies do not support the application of these measures as indices of perceived quality. For example, airflow, nodule size, and the size of a glottal gap are not good predictors of rated "breathiness."[41-43] In fact, most objective measures do not even consistently separate pathologic from normal voices.[32,43-45]

Finally, listeners contribute to these difficulties through the complexity and flexibility of the perceptual strategies they employ in perceptual tasks. The salience of a single acoustic feature may vary depending on the context in which the voice is judged, the listener's perceptual habits and preferences, momentary vagaries of attention, or many other cognitive or situational variables, even when a single voice is being judged.[46] When complex perceptual processes are combined with complex signals arising from complex vocal physiology, it is not surprising that correspondences among instrumental and perceptual measures are so variable.

To summarize, measurement of perceived voice quality is an essential component of clinical voice evaluations, because patients and their families

decide whether treatment is successful based largely on how the patient sounds, and clinicians make many decisions about managing speech and voice disorders based on perceptual judgments. Existing research does not support the substitution of instrumental measures for perceptual assessment, because instrumental measures are not meaningful in the absence of theoretical and empirical links to perception.[47] It is unfortunate, however, that existing protocols for perceptual assessment do not appear to provide reliable or valid measures of quality.

It is our belief that this problem can be resolved only by the development of new approaches to the study of voice perception. Auditory perception of voice relies in some complex fashion upon the acoustic signal that reaches the ear, and derives from the spectral envelope, noise components, their changes over time, and so on.[48] As a result, quality is inherently multidimensional,[49] and the cognitive and sensory processes involved in its auditory perception are complex and—at present—poorly understood. It should be expected that the psychoacoustics of pathological voices, which, as a class, are known for their acoustic variability, will present a considerable challenge to our understanding. Nevertheless, perception of pathologic voice signals has been studied almost exclusively with univariate (or occasionally multivariate) correlational techniques, that relate unidimensional voice quality ratings to instrumental measurements. This approach seems unlikely to prove enlightening, given that correlation does not imply causality. Simply knowing the empirical association between instrumental and perceptual variables cannot reveal how a listener exploits the acoustic signal to arrive at a perceptual response. For example, correlation coefficients do not describe a variable's perceptual salience in the context of other variables. In fact, because we have not directly studied perceptual issues such as the factors governing the relative perceptual salience of different vocal attributes, we cannot draw conclusions about the relationship between instrumental measures and perceived voice quality. In retrospect, it appears that this traditional correlational approach has been a dead-end in our attempt to understand human auditory perception of pathologic voice quality, in that these issues cannot be resolved using the experimental and statistical methods that voice research has traditionally applied in the validation of instrumental measures.

However, disenchantment with traditional approaches to the study of perception does not warrant the abandonment of perceptual assessment as part of the clinical evaluation and treatment of patients. Instead, new research approaches are needed to explicate how listeners perceive voices. Ideally, such approaches would provide concrete, cause-and-effect links between perceptual phenomena and various aspects of the acoustic signal or physiologic system (in contrast to the inferential links provided by correlational studies). For example, in analysis by synthesis,[27,28] listeners use speech synthesis to manipulate acoustic variables to construct a synthetic token that matches the quality of a natural pathologic voice of interest. The values of these acoustic parameters then directly represent a listener's perceptual response, without the need to use traditional voice quality scales of questionable validity.

In conclusion, discarding the listener in the measurement of voice quality is not in the clinician's interest. Although listeners use signals to arrive at a percept, measuring the signal alone will never provide a good index of what the listener is doing. Quality is inherently in the ear of the listener, and other measures are unavoidably indirect indices of what a listener hears. Alternate approaches to assessing listeners' perceptions of voice quality are possible and necessary, and hold promise for ongoing improvement in the clinical management of voice.

NICHOLAS SCHIAVETTI, PH.D.
DALE EVAN METZ, PH.D.

Department of Communicative Disorders and Sciences
State University of New York
Geneseo, New York

The issue to be addressed is whether or not (or under what conditions) it is possible or necessary to measure and quantify perceptions of voice quality in light of the instrumental measures of vocal characteristics that are available. As Case[50(p112)] concisely points out:

> Perceptual judgment of voice is a controversial method. Some authorities find perceptual judgment so invalid and unreliable as to be meaningless, and believe that only objective and instrumentally based measurements should be used to discriminate voices. At the opposite extreme of that position is the voice clinician who relies only on perceptual judgment of voice without any objective assessment. However, in many voice evaluations that occur daily in clinics, hospitals, private practices, and school systems, it is likely that a combination of perceptual judgment and objective assessment is the standard.

Like Case, we believe that there are conditions under which instrumental measures are appropriate, whereas other conditions may call for perceptual judgment. We would like to approach this issue, therefore, by discussing the relative advantages and disadvantages of instrumental and perceptual measures, and addressing what steps may be necessary to improve the usefulness of both instrumental and perceptual voice measures.

Instrumental Measures of Voice Characteristics

Much has been written in recent years about significant developments in electronic and digital instrumentation for the measurement of the acoustical and physiological characteristics of normal and disordered voice.[51-53] Behrman and Orlikoff[54] have reviewed the importance of the routine use of instrumental measures in clinical practice, but add that perceptual measures are also essential to the voice evaluation at all levels of the treatment regimen. They caution, however, that perceptual judgment helps define only the perceptible symptoms of dysphonia, providing no specific information on how a given speaker coordinates the respiratory, laryngeal, and articulatory subsystems or how interactions among these subsystems contribute to the perceived symptoms. Behrman and Orlikoff[54] conclude that instrumental vocal analysis may be used to establish a foundation for assessment and treatment in ways that, in all probability, cannot be achieved with perceptual measures alone. We, therefore, believe that it is important for voice clinicians to recognize the limitations of perceptual evaluations and, as Case[50(p118)] has stated, to "move toward objective assessment to complement and enhance the validity of perceptual voice judgments."

Perceptual Measures of Voice Quality

Despite the increased availability and importance of instrumental analysis, perceptual measures of voice quality are widely used and are likely to remain so for a number of reasons. For one thing, individuals tend to seek treatment because of the perception that they (or someone in their environment) have about their voice quality. Thus, those seeking treatment are doing so to "hear" improvement and perceptual feedback can be used to guide the direction of treatment.

Perceptual measures of voice quality have *face validity* in the sense that Anastasi[55(p144)] defines the term:

> face validity . . . is not validity in the technical sense; it refers, not to what the test actually measures, but to what it appears superficially to measure. Face validity pertains to whether the test 'looks valid' to the examinees who take it, the administrative personnel who decide on its use, and other untrained observers. Fundamentally, the question of face validity concerns rapport and public relations. Nonetheless Anastasi[55] emphasized the desirability of face validity, concluding that "if test content appears irrelevant, inappropriate, silly, or childish, the result will be poor cooperation, regardless of the actual validity of the test.

But, while commonly used, are perceptual measures of voice quality *useful*? As for any measurement, their value depends both on reliability and validity. As Onslow and Ingham[56(p5)] stated:

> In some respects, the pattern of research on voice disorders or voice quality seems to imply that perceptual rating scales are the 'poor cousins' of, ostensibly, more precise and more valid objective measures. The limited reliability and uncertain validity of some of the rating scale labels contrasts strikingly with the technology-based measurement approaches. However, this unflattering comparison cannot be used to bypass the fact that, in the final analysis, any attempt to measure or modify voice quality must be validated by listeners, or in some cases, speakers. In short, subjective judgment systems will probably continue to be validity checks, at least, in this area of therapy management. That role will undoubtedly be more prominent when the validity and reliability of these systems are refined and improved.

Many of the specific reliability and validity issues that relate to improving the usefulness of perceptual measures have been discussed in depth in both the behavioral research methodology and speech pathology literature. We will address a few of the most salient points below.

Reliability

Cordes[57] has said that the "general trustworthiness of obtained data" or the "dependability, consistency, predictability, and stability" are common reliability concerns in communicative disorders when researchers consider the question of whether or not the observed "data could be reproduced if the same subjects were tested again under similar circumstances." Reliability is also tied to measurement precision.[58] The "true score" model of measurement accuracy examines the ratio of the true-score variance to the observed-score variance. In theory, the true score may be regarded as the score that would have been obtained if the conditions of measurement were ideal. Since ideal conditions are never possible, the score that is actually obtained

(the observed score) is always subject to a certain amount of error.

Steps taken to isolate and reduce the error components of observed scores, then, help to improve the reliability of measurements. Kreiman and her colleagues[46] have proposed a theoretical framework for identifying sources of variability in voice quality measures (such as rating task, listener biases, etc.) that may prove useful in the identification and reduction of error components. It should be pointed out, however, that there has been a great deal of confusion in the literature regarding the terms "reliability" and "agreement." According to Kreiman et al.[46(p22)]:

> Listeners are in agreement to the extent that they make exactly the same judgments about the voices rated. Ratings are reliable when the relationship of one rated voice to another is constant (ie, when voice ratings are parallel or correlated), although the absolute rating may differ from listener to listener. The number of ratings that agree exactly or within ± one scale value is a common measure of agreement; the intraclass correlation is one measure of rating reliability.

Cordes[57] has further distinguished between inter- and intra-observer agreement, the degree of correspondence among the scores assigned by different raters, or by the same rater on different occasions, respectively.

Much can be done to improve reliability and agreement between and among raters. A recent study by Gerratt et al,[59] for example, found that greater reliability and agreement were obtained when an external standard was used to anchor judgments of "vocal roughness" in synthetic stimuli. Thus the development of methods that allow raters to match perceived voice characteristics to external (explicit and constant) rather than internal (unobservable and unstable) standards hold promise for substantially improving the reliability and agreement in assigning scale values. Another concern, of course, is the validity of such values.

Validity

When we ask about the validity of a testing procedure, we are inquiring as to how closely the test measures what it is that we want to measure, if at all. For instance, if we wish to measure a specific voice quality, such as "vocal roughness," we do not want other voice or speech characteristics (such as "hypernasality") to influence the measurement.

Thorndike and Hagen[60] discuss in detail three types of validity:

1. *content validity*, the degree to which a measure is representative of the behavior it purports to measure;
2. *criterion validity*, the degree to which a measure correlates with another known measure of the behavior it purports to measure; and
3. *construct validity*, the degree to which a measure reflects a theoretical construct.

The validity of perceptual voice quality measures have been addressed in a number of ways, including correlating perceptual ratings with acoustic and aerodynamic measures, evaluating the nature of the labels used to describe voice qualities, determining the extent to which voice qualities differentiate normal from abnormal voices, specifying the criteria used to define "normal" voice quality, surveying the manner in which different voice qualities vary among voice disorders, and assessing the relationship between perceived voice quality and the severity of vocal disorder.[56] Nonetheless, a great deal of research remains to be done to establish the validity of perceptual voice quality measures.

As useful as instrumental measures may be, they are not likely to displace perceptual measures of voice quality. Rather, it is more likely that instrumental and perceptual measures will be used to complement each other. Ultimately, the degree of clinical utility of both instrumental and perceptual measures will be based on the degree to which both types of measures are reliable and valid—"qualities," according to Thorndike and Hagen,[60] "desired in any measurement procedure."

JAMES DEMBOWSKI, PH.D.
BEN C. WATSON, PH.D.

State University of New York—New Paltz
New Paltz, New York
New York Medical College
Valhalla, New York

Is it possible or necessary to quantify perceptions of voice quality and, if so, for what purposes? The use of perceptual impressions in the identification, diagnosis, and treatment of voice disorders raises question related to the difficulty of mapping these impressions against measurable physical phenomena. How might nominal perceptual categories be defined relative to the continuous physiologic processes that form the bases of the perception? How might judgments be scaled within a perceptual category, or

used for differential diagnosis, when no isomorphic relation exists between perceptual impressions and the underlying physiologic processes?

We address these questions in terms of some basic assumptions underlying two recent studies.[46,59] Particularly with respect to the issue of whether multiple listeners can reliably scale their perceptual impression of disorder across multiple talkers, we question the clinical need for such reliability. We propose that definitional issues pose a more pressing problem. As a convenient example, consider the rating of the perceptual impression of "roughness," explored in these studies. The authors focus on the difficulty of obtaining reliable inter-rater judgments of degrees of vocal "roughness." However, the ability to judge the degree to which a voice exhibits a given quality is predicated upon being first able to define the thing being judged. What constitutes a "rough" voice? The term is one largely absent from at least one commonly used clinical text on voice disorders,[61] which does, however, frequently refer to a "hoarse" voice quality. Are the two terms synonymous? Even if the answer is yes, then what constitutes "hoarseness?" What are the appropriate terms for defining such perceptual impressions? We believe that, as with many other perceptual categories, definitions should be derived, to the extent that they can be, from observable, measurable, physical phenomena. That is, perceptual impressions need to be defined according to the acoustic and physiologic processes that give rise to the impressions.

To draw an analogy from well beyond the field, consider how one might define, let's say, a chocolate brownie. And how is a "brownie" distinct from, say, a chocolate sponge cake or a torte? One defines such a thing first with reference to its ingredients, its measurable, physical ingredients, and the physical processes by which they are combined and baked. You could say that the brownie is defined by its anatomy and physiology. Change the anatomy (alter the number of eggs, the proportion of flour, the type of chocolate), or change the physiology (alter the baking time, the order in which the ingredients are combined, or the chemical leavening agents) and you get, at least, some different sort of brownie than you started with, and maybe something not a brownie at all. This analogy is not intended to suggest a frivolous attitude toward perceptual judgments. Rather, the analogy suggests how, even in relatively trivial matters of daily life, we need to define a thing in terms of components that are at least observable, and preferably measurable, if we want to reliably identify, reproduce, or evaluate that thing (be it brownie or voice quality). Likewise, "vocal roughness" must arise from some physical traits that, ideally, would be as easily identified and measured as flour and eggs: say, a given signal-to-noise ratio, abnormal degree of airflow, irregularity in period of vocal fold vibration, and fundamental frequency (to mention just a few possibilities). In fact, however, our field has not yet clearly identified and quantified the physical components of vocal qualities such as "roughness," "hoarseness," "breathiness," "stridency," and a host of other impressions that have found their way into clinical use. For this reason we believe that an issue even more pressing than the reliable rating of such qualities is their definition in observable, measurable physical terms.

Even if a perceptual impression is clearly defined in physical terms, judging the degree to which a vocal quality is present along some scale raises additional problems. One problem is that most scaling procedures involve classifying continuously varying physical traits into discrete ordered categories—whether a clinician classifies a voice into ordinal categories from "mildly rough" to "very rough," or a more rigorous magnitude estimation procedure is employed. That is, there generally will be a range of voice qualities and observable physical traits for any given point on a scale of judgments. In the baking analogy, there may be several brownie recipes that produce more-or-less "fudgy" (or, conversely, "cake-like") brownies, or there may be multiple recipes that fall into the "mighty fine" (or "feed-these-to-the-dog") category. Gerratt, Kreiman, and their colleagues[46,59] addressed this issue in part through the use of visual analog scales that are undifferentiated lines, and therefore continuous like the variable being judged. But even the use of a continuous scale for judging a continuous percept fails to address the problem that a perceptual impression likely arises from several simultaneous physical traits, all of which vary continuously. Furthermore, these traits may combine in different ways to produce similar perceptual impressions. For example, the perception of "roughness" may result from some combination of high glottal airflow, irregularity in fundamental frequency, and unexpectedly low fundamental frequency. Differently weighted contributions of these factors may produce comparable (even identical) perceptions of "roughness." In short, the relationship between perceived voice qualities (or magnitudes of a given quality) and physical phenomena that underlie those qualities, is a one-to-many relationship. Any number of acoustic phenomena, and underlying those acoustic phenomena, any number of physiological factors, may contribute to a perception of a "rough" voice.

The observations that a range of voice qualities may map against any single discrete step on an evaluation scale, that multiple physical traits may contribute to a categorical perception of voice quality, and that the contributing physical traits might covary in imperceptible ways, all provide examples of the ways in which perceptual categories are not isomorphic with physiologic phenomena. The notion that there is no simple one-to-one mapping of perceptual category to the acoustic and physiologic features of that category has clinical implications. Specifically, the anisomorphism of perceptual category and physiology renders the task of identifying different vocal pathologies on the basis of what a clinician hears difficult, if not impossible. This, in turn, suggests that the scaling of perceived vocal impressions may have limited clinical application, even if it could be done reliably. That is, the quantification of any perceptual category in and of itself, without reference to (and, preferably, quantification of) the factors that give rise to the signal perceived by the listener, is of limited utility in the evaluation, differential diagnosis, and treatment of voice disorders.

It is not uncommon for multiple pathologies to produce similar, even identical, acoustic-perceptual vocal qualities. For this reason the extent to which the nature of a problem can be determined solely by perceptual evaluation is limited. "Hoarseness," for example, is reported to be characteristic of many vocal pathologies of vastly different etiologies—and of vastly different consequences for the health of an individual. For this reason, we question the assumption of Gerratt et al[59(p14)] that the "perception of a patient's voice is at the heart of evaluating and treating patients with voice disorders." Far from being at the heart of evaluating and treating patients, in our opinion, the perception of voice is only the most superficial and preliminary step toward determining whether a patient has a problem, and tells little about the nature of the problem.

Fortunately, the ability to examine the nature of a voice problem, that is, to analyze the physiological "ingredients" of vocal pathology, is potentially more available to voice clinicians than to many others who work in speech-language pathology. Even if many details remain opaque, the fundamental relationship between perceivable vocal function and laryngeal physiology is largely indisputable (compared with many developmental and neurological pathologies where the relationship between function and physiology is less easily observed and less well established). The etiology of most voice disorders is functional and is observable in the peripheral physiology. In some cases even those voice disorders associated with CNS dysfunction may produce consistent symptoms at the observable periphery (the larynx) that are amenable to peripherally directed treatment (such as visual biofeedback of degree of vocal fold approximation). Voice clinicians are also fortunate (compared, say, with their colleagues in articulation) in that there are several clinically practical technologies for examining vocal physiology (stroboscopy, EEG, air pressure and flow devices, etc.). Despite the potential availability of such technology, it has been reported that many clinicians continue to prefer perceptual evaluations over instrumental measures. However, the frequent use of a measure is no argument for its validity, or its superiority to other measures. In both clinic and lab, some measures may be preferred because they are what the examiners happen to know, because they are relatively simple, because they are cheap, or because the equipment or expertise to make different measures is unavailable. Simple practicality and precedent, not to mention fear of the unknown, may exert a powerful influence on both clinical and research procedures. Examiners may employ certain measures simply because they can, and avoid others that may be technically complex or time consuming even if they are potentially more valid, reliable, precise, or revealing of a disorder's causes. Certainly perceptual judgments, that require little or no technical equipment (aside, perhaps, from a tape recorder), are relatively quick and easy to make. In this sense, they are practical. But the time- and effort-saving practicality of perceptual judgments may be offset by the potential for error, and the likely need in any event for supplementary instrumental measures to provide reliable differential diagnoses. No doubt some clinicians with the right degree of education and experience have honed their ears well enough that they can draw some useful diagnostic conclusions from what they hear. However, it would be a Sisyphean task to try to codify just what type and degree of education and experience produces a clinician who validly and reliably generates diagnostic conclusions based on perception that are consistent with the conclusions of other similarly educated and experienced professionals as well as with the inferences that may be drawn from direct and indirect instrumental measures of physiology. On the other hand, given the current state of technology and voice science, there may be little need for such inter-evaluator reliability in perceptual voice judgments. That is, because the relationship between peripheral physiology and vocal function is relatively well understood, and because the vocal mechanism is

relatively easily observed, there may be relatively little clinical need for even reliable perceptual judgments of voice quality.

This is not to argue that there is no place in the clinic for perceptual impressions. They have value in two respects: (1) as a preliminary means of identifying the existence of a problem, and (2) as a means of tracking functional improvement (or deterioration) in the voice over time. However, for neither of these purposes do listeners need to reliably scale voice quality judgments *across* speakers. As a preliminary indication of the existence of a voice problem, a perceived "poor" or "abnormal" voice quality may prompt a patient to seek advice. But, importantly, this is chiefly a judgment on the part of the speaker (and the speaker's close acquaintances) and is a judgment formed with respect to the patient's knowledge of his own normal voice. When the speaker perceives a change from his perceptual standard of his own normal voice, he may consider that symptomatic of a problem needing attention. Thus a preliminary indication of voice problem might be a speaker's intra-rater judgment of a change in voice quality. A clinician, however, is unlikely to have the same standard of normality as the speaker (not having heard the speaker before an initial consultation), and is therefore in a poor position to judge even the existence of a problem without extra-perceptual measures, except in extreme cases. Thus, the perception of abnormality and diagnosis of disorder (prior to confirmation of some disease process or trauma) depends on some standard of normality, not across speakers, but for the speaker under consideration. Additionally, the perception of disorder must be related to the functional value of the patient's voice. Anecdotally speaking, we can all think of certain popular performers who possess(ed) what might be considered, by some standards, "pathological" voices, but whose voices were functional for those people, perhaps even vital for maintaining the image that contributed to their livelihood (the singers Louis Armstrong, Marianne Faithful, and Tom Waits, and the actress Lauren Bacall, come immediately to mind, though there are certainly many other examples). Conversely, consider the anecdote of a patient known to us who suffered laryngeal trauma in an accident and required surgical reconstruction of her larynx. Post-operatively, doctors and voice clinicians judged the patient's voice as functional, and even attractive. The patient, however, found it abnormal because it was not the same as her pre-traumatic voice (which her medical caretakers had never heard). As it was unlikely that the patient would ever be able to recover her pretraumatic voice, therapy goals were in part a matter of teaching the patient that the voice of her reconstructed larynx was functional for her needs, in effect recalibrating her standard of normality.

A second use for perceptual judgments is that they may provide some general indication of therapeutic progress, particularly to the degree that progress needs to be demonstrated to non-professionals: the patient, the patient's family, or insurance representatives. Again, however, therapeutic progress (or degeneration, when tracking effects of some disease process) is best measured with reference to a patient's own previous voice. This requires reliably judging change in a patient's voice, but does not require reliable agreement among clinicians across speakers. If Therapist Sue judges a patient's voice as "severely rough" before therapy and "mildly rough" after, while Therapist Henry judges the same voice as "moderately rough" before therapy and "within normal limits" after therapy, everyone agrees that the goal of improvement through therapy has been met. The perception of change in this case is reliable across evaluators, even if the perceptual impression of specific scale values is not. We would argue that the reliable perception of change is of greater clinical importance than reliable scaling of a voice quality at a single point in time. That is, the type of reliability likely to be of greatest clinical value is intra-rater (clinician) reliability in the judgment of individual voices, rather than inter-rater reliability in judgments across speakers.

There might be substantial clinical value in seeing the methodological rigor of Kreiman's and Gerratt's work[46,59] applied to studies of intra-rater perception of individual voices. Their finding that voice ratings in the context of "anchored" comparisons were more reliable than non-comparative judgments suggests that listener reliability in other comparative voice judgments might be quite good. An interesting research goal might be to attempt to develop a description of "just noticeable differences" in the perception of individual's voice qualities, or, better still, to explore the JNDs in the separate acoustic components (signal to noise ratio, fundamental frequency, jitter, etc.) of a perceived voice quality.

In brief, if we need to know whether this week's batch of brownies is an improvement over last week's, we need only have a reliable comparison of the two batches of brownies. If this week's dessert is still considered best for beasts, we likely gain little by worrying over the extent to which our friends agree on just *how* inedible the brownies are. We would more likely produce a better brownie next week by looking to the "anatomy" of our recipe or recalibrating the "physiology" of our oven.

JOHN A. HASKELL, ED.D.

Speech Pathologist
New York, New York

The assessment and treatment of voice disorders is both an art and a science. Both are critical to the appropriate gathering of information, to the meaningful interpretation of results, and to the choices made in the management of the individual patient. The clinician's wisdom in making choices is based on many things: objective data yielded in formal testing, perceptual judgments (both auditory and visual), knowledge gleaned from talking with the patient, the patient's vocal self-perception,[62] the clinician's knowledge about the nature of vocal disorders, and, ultimately, the clinician's experience—both professional and personal. The art of clinical practice is how they are put together and, in the long run, this synthesis determines our effectiveness as clinicians. The question of whether there is a clinical need to measure listener-perceived voice quality, with the availability of information from instrumentation, may be viewed as a matter of art versus science or, more appropriately, as a balance of art *and* science.

Most voice clinicians would most likely agree that the information from our present (and ever-developing) capability to objectively assess voice characteristics has become an essential part of the diagnostic process. Perceptual judgments, however, form another essential part. We continue to judge what we hear and see because, first, it is natural to do so; second, perceptual impressions frequently provide information that contrasts with objective data; and third, frequently there is a need to compare judgments with those of our patients[63] and with other professionals involved in the patients' care. If anything, objective physical data help clinicians refine their judgments, so that types of "roughness," for example, can be discriminated usefully.

A concern of mine is that clinicians may move away from perceptual judgments, not because of the availability of sophisticated instrumentation, but because of three forces that have dramatically altered clinical conduct, namely the use of computer programs for report writing, time pressure from managed health care, and reimbursement-driven accountability. In general, these forces are pressuring clinicians, not only to derive "numbers," but to do so quickly. As a consequence, clinicians may be giving less time—and perhaps less thought—to the synthesis of information, making service delivery far less customized to the individual patient. In order to derive a quick and reasonable diagnosis and treatment strategy, then, much of the art and science of the clinical process may be sacrificed.

I am also concerned about the training of students who will be treating patients with voice disorders. For the most part, students who are currently completing their professional training have grown up with computers and have developed what might be called a "computer mentality." Many have been fortunate to have had the opportunity to learn about voice and voice disorders through hands-on experience with stroboscopy, airflow and air pressure devices, electroglottography, real-time acoustic analysis systems, among others. But are they being taught to listen? Perhaps more than ever before it has become critical to stress the importance of using instrumentation to complement perceptual judgments of voice quality in the assessment and treatment of voice disorders.

But how do we teach our students to listen? And how *do* we integrate perceptual judgments with instrumental measures? Does quantifying the clinician's perceptual impressions accomplish this purpose? Although no universally accepted rating scale exists and there is little agreement with respect to descriptive terminology, many clinics worldwide have included various types of perceptual "indices" on their standard diagnostic forms. For instance, judgments of pitch, voice range, or quality ("roughness," "breathiness," "strain," etc.) may be rated by the clinician or the patient may be asked to rate his or her judgment of severity, discomfort, effort, or impact using a multi-point scale of some sort. Such rating scales are useful, but there are potential dangers in this type of evaluation. While these kinds of forms may serve some patients well, they may be inadequate for others—perhaps for those whose pathology is particularly unclear and whose abnormal vocal behaviors are less consistent and predictable.

The use of listener or patient rating scales does not replace objective data, as it is well recognized by any experienced clinician that perceptual judgments do no necessarily correspond to the severity of the laryngeal or vocal disorder. If physical data can differentiate voice qualities and document underlying laryngeal and vocal function, then perhaps they should be used to establish valid perceptual scales with improved reliability. If nothing else, instrumentation should be used to help the clinician listen with greater focus and to understand, in a more sophisticated and detailed way, a patient's variant and deviant vocal behaviors. In this way, we may be able to counteract the threat of superficial management of patients while maintaining a sensible balance of art and science in the process.

A FEW CONCLUDING REMARKS

Although there is often considerable disagreement, in my opinion, each of the roundtable contributions raises valid and noteworthy points and presents well-reasoned arguments in their support. Such differing points of view are to be expected perhaps, not simply because the issues addressed are both complex and multifaceted, but because they relate quite directly to the basic underlying principles that a clinician applies to guide and direct his or her decision making. It is not my intention to further debate the individual points that were raised *per se*, but rather to attempt to summarize the arguments that were advanced while placing the major issues within a more global clinical perspective. Despite the variety of opinions and concerns, I think that it could be said that, to varying degrees, the roundtable participants were drawn toward one of two viewpoints.

At one extreme is the view that, because auditory perceptions form the basis of a patient's complaints, perceptual judgments should take a primary role in guiding vocal assessment and management. As such, effective scaling of auditory vocal impressions can only serve to make the diagnostic and therapeutic processes more exact and systematic. Objective physical measures, on the other hand, may be helpful in supporting the clinician's perceptual judgment, but they may also needlessly *divert attention from the patient's symptoms* and the primary reason the patient has sought professional help.

At the other extreme is the notion that perceptions of voice quality do not provide an awareness of the physiologic details that give rise to the acoustic product nor an appreciation of the acoustic details that have combined to give rise to a given perception. Furthermore, perceptual judgments are inherently biased by listener characteristics and are thus far from immutable. Therefore, rather than listener-based rating scales, objective physical measures are of primary importance to the voice clinician since they necessarily *call attention to the cause of the patient's perceived symptoms*. Given that it is the clinician's responsibility to effect change, the physiologic and acoustic abnormalities must first be identified and characterized so that they may be specifically addressed and systematically monitored.

How, then, might these views which, at least at first seem to be mutually exclusive, nonetheless be reconciled? Clearly the auditory perceptions that form the basis of a patient's complaints cannot be ignored by the clinician. However, the problems and pitfalls associated with the documentation of those impressions are foretold by the difficulty the patients often express as they themselves attempt to describe the quality of their voices. When the plethora of terms that are commonly employed in the description of voices (both good and bad, normal and abnormal) are surveyed, it is perhaps worthy of note that relatively few of them relate specifically to sound or audition. For instance, voices may be described in visual (eg, *clear, brilliant, bright, dark*), gustatory (eg, *dulcet, mellow, sweet, mellifluous, syrupy*), or kinesthetic terms (eg, *stressed, strained, strangled, pressed, forced, grating, firm, tight, soft, wobbly, tense, gargled, abrasive, raspy*). While a voice might be described using an anatomical reference (eg, *nasal, throaty, guttural, chest, head, pectoral*), it can also be likened to an inanimate object (eg, *metallic, brassy, wooden, woolly, muddy, velvety, silken, gravelly*), a texture (eg, *rough, coarse, dull, sharp, smooth, choppy*), or some material characteristic (eg, *flat, thin, husky, hardy, weak, fragile, robust, light, heavy*). Even labels that are tied almost exclusively to the "sound of the voice" are often difficult to define with acceptable reliability and precision. It has been my experience, for instance, that there are many professionals, especially those in the medical setting, who use the terms "hoarseness" and "dysphonia" interchangeably. What, then, do these perceptually based appellations add and how might their scaling benefit the clinical process?

Personally, I do feel that subjective assessments of perceived voice quality serve an important clinical purpose. It is crucial to keep in mind, however, that though patients may pursue vocal assessment because they recognize deficiencies in the perceived "sound" of their voices, their intent is not to solicit the clinician's aesthetic appraisal of their vocal symptoms. A patient, by and large, is not looking to have her own auditory impressions confirmed or refuted and, by extension, is not seeking a professional opinion of whether, for instance, her voice would be best categorized as "rough," rather than "harsh," "breathy," or "hoarse." To be sure, patients seek the assistance of the voice clinician in an effort to make themselves "sound better" and thus to communicate more effectively. The purpose of auditory-perceptual judgments, then, is to help define *long-term* goals. Any attempt to scale, standardize, or in any way quantify those perceptions should thus be done with an eye toward facilitating the assessment of therapeutic outcome. Subjective auditory impressions, especially those of the patients themselves, provide the voice clinician with important information about vocal disability, ineffectiveness, and handicap. A patient, for instance, would find little consolation in knowing that his mean vocal jitter is within normal limits following a course of therapy if his perceived vocal symptoms

continue. Since verbal communication depends on both the speaker and the listener, perceptual judgments are important considerations in determining the ultimate success or failure of therapy.

Frankly, objective physical measures are ill suited to assess the disability and handicap speakers face as a consequence of their voice disorder. In my opinion, efforts to relate acoustic measures to perceived voice characteristics are largely misguided. It is implausible that any one acoustic measure (or "collection" of measures) will serve as an index of any aesthetic category of voice quality since, unlike the sound pressure signal, such judgments depend critically on the listening context, the linguistic content of the utterance, and the vagaries of the nature and nurture of the listener. There is no acoustic parameter—no matter how complex—that can serve as a measure of "hoarseness," any more than fundamental frequency can be used as a direct measure of the perceived "pitch of the voice." Rather, acoustic measures are valuable only to the extent that they allow valid and reliable inferences to be made about what behaviors need to be modified.[54,64] Several of the roundtable contributors cited the shortcomings of jitter measures. I, too, recognize the fact that perturbation measures, in general, have largely failed to live up to the clinical promise they once held. This is not due to the rather tenuous relationship they have to perceived "hoarseness," "roughness," or "harshness," but rather to the fact that they rarely provide unequivocal information about laryngeal or vocal function. In this vein, it has long been my experience that the clinical utility of any acoustic assessment is enhanced when the speech product can be compared with at least one simultaneously acquired physiologic signal, be it glottographic, endoscopic, or aerodynamic.

Unless it is the clinician's intent to alter the way a patient perceives his or her voice or to provide an external communication aid such as a voice amplifier, reducing vocal disability and associated handicap is achieved by improving vocal behavior. One of the fundamental purposes of any voice assessment is to establish precisely which behavioral changes are needed. In ways in which auditory-perceptual judgments are ill suited, objective voice measures help identify the impairments in vocal physiology that result in the patient's vocal disorder. The clinical import of acoustic and physiologic measures is thus tied to their ability to help outline, guide, and monitor the *short-term* goals of therapy. It is this, I believe, which is at the heart of efforts to improve clinical accountability by means of objective measurement. Attention to vocal physiology also serves to address symptoms that are unrelated to sound quality, such as perceived vocal effort, fatigue, and discomfort. A voice that is "acceptable" to listeners, after all, may be produced with an inappropriately high driving pressure, low glottal resistance, or excessive vocal-fold compression. Attention to such details helps the clinician meet another clinical responsibility: modifying behaviors that predispose the patient to develop future voice symptoms.

It is unreasonable to suppose that this roundtable will resolve the controversy surrounding the issue of auditory-perceptual judgment in the assessment and management of dysphonia. Nonetheless, it is hoped that clinicians will be encouraged to contemplate the arguments and opinions provided as they relate to their own clinical practice. As theory, technique, and technology improves, such issues will almost assuredly move to the forefront of clinical debate.

Acknowledgments

Preparation of B. R. Gerratt and J. Kreiman's contribution to this roundtable discussion was supported by NIDCD grant number 01797. Drs. Gerratt and Kreiman thank Kristin Precoda, Andrew Erman, and Melissa Epstein for valuable comments on an earlier version of their discussion. Note: Individual contributions from roundtable participants may have been edited for content, style, and length.

Address correspondence to

Robert F. Orlikoff, Ph.D.
Memorial Sloan-Kettering Cancer Center
1275 York Ave., Box 403
New York, NY 10021
e-mail: orlikofr@mskcc.org

References

1. Jensen PJ. Adequacy of terminology for clinical judgment of voice quality deviation. *Eye Ear Nose Throat Monthly.* 1965;44:77-82.
2. Sonninen A. Phoniatric viewpoint on hoarseness. *Acta Otolaryngol.* 1970;263:68-81.
3. Blaustein S, Bar A. Reliability of perceptual voice assessment. *J Commun Disord.* 1983;16:157-161.
4. Bassich CJ, Ludlow CL. The use of perceptual methods by new clinicians for assessing voice quality. *J Speech Hear Disord.* 1986;51:125-133.
5. Kearns KP, Simmons NN. Interobserver reliability and perceptual ratings: More than meets the ear. *J Speech Hear Res.* 1988;31:131-136.
6. Neiman GS, Applegate JA. Accuracy of listener judgments of perceived age relative to chronological age in adults. *Folia Phoniatr.* 1990;42:327-330.

7. Fex S. Perceptual evaluation. *J Voice*. 1992;6:155-158.
8. de Krom G. Consistency and reliability of voice quality ratings for different types of speech fragments. *J Speech Hear Res*. 1994;37:985-1000.
9. Kent RD. Hearing and believing: Some limits to the auditory-perceptual assessment of speech and voice disorders. *Am J Speech Lang Pathol*. 1996;5(3):7-23.
10. Gelfer MP, Young, SR. Comparisons of intensity measures and their stability in male and female speakers. *J Voice*. 1997;11:178-186.
11. Gelfer MP. Fundamental frequency, intensity, and vowel selection: Effects on measures of phonatory stability. *J Speech Hear Res*. 1995;38:1189-1198.
12. Titze IR, Winholtz WS. (1993). Effect of microphone type and placement on voice perturbation measurements. *J Speech Hear Res*. 1993;36:1177-1190.
13. Jiang J, Lin E, Hanson DG. Effect of tape recording on perturbation measures. *J Speech Hear Res*. 1998;41:1031-1041.
14. Titze IR, Liang H. Comparison of F_0 extraction methods for high-precision voice perturbation measurements. *J Speech Hear Res*. 1993;36:1120-1133.
15. Hirano M. *Clinical Examination of Voice*. New York: Springer-Verlag; 1981.
16. Dejonckere PH, Obbens C, de Moor GM, Wieneke GH. Perceptual evaluation of dysphonia: Reliability and relevance. *Folia Phoniatr*. 1993;45:76-83.
17. Muller R. Heiserkeit. *Ther Umsch*. 1995;52:759-762.
18. Dejonckere PH, Remacle M, Fresnel-Elbaz E, Woisard V, Crevier-Buchman L, Millet B. Differentiated perceptual evaluation of pathological voice quality: Reliability and correlations with acoustic measurements. *Rev Laryngol Otol Rhinol*. 1996;117:219-224.
19. Dejonckere PH, Lebacq J. Acoustic, perceptual, aerodynamic and anatomical correlations in voice pathology. *ORL*. 1996;58:326-332.
20. Dejonckere PH. Cepstral voice analysis: Link with perception and stroboscopy. In: *Sydney '97: XVI World Congress of Otorhinolaryngology—Head and Neck Surgery*. Bologna, Italy: Monduzzi Editore; 1997;1661-1665.
21. Millet B, Dejonckere PH. What determines the differences in perceptual rating of dysphonia between experienced raters? *Folia Phoniatr Logoped*. 1998;50:305-310.
22. Eskenazi L, Childers DG, Hicks DM. Acoustic correlates of vocal quality. *J Speech Hear Res*. 1990;33:298-306.
23. Martin D, Fitch J, Wolfe V. Pathologic voice type and the acoustic prediction of severity. *J Speech Hear Res*. 1995;38:765-771.
24. de Krom G. Some spectral correlates of pathological breathy and rough voice quality for different types of vowel fragments. *J Speech Hear Res*. 1995;38:794-811.
25. Takahashi H, Koike Y. Some perceptual dimensions and acoustical correlates of pathologic voices. *Acta Otolaryngol*. 1975;268 (suppl. 338):1-24.
26. Wolfe V, Fitch J, Martin D. Acoustic measures of dysphonic severity across and within voice types. *Folia Phoniatr Logoped*. 1997;49:292-299.
27. Kreiman J, Gerratt BR. Validity of rating scale measures of voice quality. *J Acoust Soc Am*. 1998;104:1598-1608.
28. Kreiman J, Gerratt BR. The perceptual structure of pathologic voice quality. *J Acoust Soc Am*. 1996;100:1787-1795.
29. Denes PB, Pinson EN. *The Speech Chain*. 2nd ed. New York: Freeman; 1992.
30. Titze IR, Horii Y, Scherer RC. Some technical considerations in voice perturbation measurements. *J Speech Hear Res*. 1987;30:252-260.
31. Arends N, Povel D, van Os E, Speth L. Predicting voice quality of deaf speakers on the basis of glottal characteristics. *J Speech Hear Res*. 1990;33:116-122.
32. Heiberger VL, Horii Y. Jitter and shimmer in sustained phonation. In: Lass NJ, ed. *Speech and Language: Advances in Basic Research and Practice*. Vol. 7. New York: Academic Press; 1982:299-332.
33. Ludlow CL, Bassich C, Connor NP, Coulter DC, Lee YJ. The validity of using phonatory jitter and shimmer to detect laryngeal pathology. In: Baer T, Sasaki C, Harris KS, eds. *Laryngeal Function in Phonation and Respiration*. Boston, MA: College Hill Press; 1987:292-508.
34. Hammarberg B, Fritzell B, Gauffin J, Sundberg J, Wedin L. Perceptual and acoustic correlates of abnormal voice qualities. *Acta Otolaryngol*. 1980;90:441-451.
35. Klatt DH, Klatt LC. Analysis, synthesis, and perception of voice quality. *J Acoust Soc Am*. 1990;87:820-857.
36. Kreiman J, Gerratt BR, Precoda K. Listener experience and perception of voice quality. *J Speech Hear Res*. 1990;33 103-115.
37. Prosek RA, Montgomery AA, Walden BE, Hawkins DB. An evaluation of residue features as correlates of voice disorders. *J Comm Disord*. 1987;20:105-117.
38. Hirano M, Hibi S, Yoshida T, Hirade Y, Kasuya H, Kikuchi Y. Acoustic analysis of pathological voice. *Acta Otolaryngol*. 1988;105:432-438.
39. Klich R. Relationships of vowel characteristics to listener ratings of breathiness. *J Speech Hear Res*. 1982;25:574-580.
40. Kreiman J, Gerratt BR. Measuring voice quality. In: Kent R, Ball MJ, eds. *Handbook of Voice Quality Measurement*. San Diego, CA: Singular; 1999:in press.
41. Rammage LA, Peppard R, Bless DM. Aerodynamic, laryngoscopic, and perceptual-acoustic characteristics in dysphonic females with posterior glottal chinks: A retrospective study. *J Voice*. 1992;6:64-78.
42. Murry T, Singh S, Sargent M. Multidimensional classification of abnormal voice qualities. *J Acoust Soc Am*. 1977;61:1630-1635.
43. Hertegård S, Gauffin J. Insufficient vocal fold closure as studied by inverse filtering. In: Gauffin J, Hammarberg B, eds. *Vocal Fold Physiology: Acoustic, Perceptual, and Physiological Aspects of Voice Mechanisms*. San Diego, Calif: Singular Publishing Group; 1991:243-250.
44. Klingholz F, Martin F. Speech wave aperiodicities at sustained phonation in functional dysphonia. *Folia Phoniatr*. 1983;35:322-327.

45. Hecker MHL, Kreul EJ. Descriptions of the speech of patients with cancer of the vocal folds. Part I: Measures of fundamental frequency. *J Acoust Soc Am.* 1971;49:1275-1282.
46. Kreiman J, Gerratt BR, Kempster GB, Erman A, Berke GS. Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research. *J Speech Hear Res.* 1993;36:21-40.
47. Catford JC. *Fundamental Problems in Phonetics.* Bloomington, IN: Indiana University Press; 1977.
48. Plomp R. *Aspects of Tone Sensation.* London, UK: Academic Press; 1976.
49. Kreiman J, Gerratt BR, Berke GS. The multidimensional nature of pathologic vocal quality. *J Acoust Soc Am.* 1994;96:1291-1302.
50. Case JL. *Clinical Management of Voice Disorders.* Austin, TX: Pro-ed; 1996.
51. Orlikoff RF, Baken RJ. *Clinical Speech and Voice Measurement: Laboratory Exercises.* San Diego, Calif: Singular Publishing Group; 1993.
52. Titze IR. *Principles of Voice Production.* Englewood Cliffs, NJ: Prentice-Hall; 1994.
53. Baken RJ, Orlikoff RF. *Clinical Measurement of Speech and Voice.* 2nd ed. San Diego, Calif: Singular Publishing Group; 1999.
54. Behrman A, Orlikoff RF. Instrumentation in voice assessment and treatment: What's the use? *Am J Speech Lang Pathol.* 1997;6(4):9-15.
55. Anastasi A. *Psychological Testing.* 4th ed. New York: MacMillan; 1988.
56. Onslow M, Ingham RJ. Speech quality measurement and the management of stuttering. *J Speech Hear Disord.* 1987;52:2-17.
57. Cordes AK. The reliability of observational data: I. Theories and methods for speech-language pathology. *J Speech Hear Res.* 1994;37:264-278.
58. Pedhazur EJ, Schmelkin LP. *Measurement, Design, and Analysis: An Integrated Approach.* Hillsdale, NJ: Lawrence Erlbaum; 1991.
59. Gerratt BR, Kreiman J, Antonanzas-Barroso N, Berke GS. Comparing internal and external standards in voice quality judgments. *J Speech Hear Res.* 1993;36:14-20.
60. Thorndike RL, Hagen E. *Measurement and Evaluation in Psychology and Education.* 3rd ed. New York: Wiley; 1969.
61. Colton RH, Casper JK. *Undersanding Voice Problems: A Physiological Perspective for Diagnosis and Treatment.* 2nd ed. Baltimore, MD: Williams and Wilkins; 1996.
62. Haskell JA. Vocal self-perception: The other side of the equation. *J Voice.* 1987;1:172-179.
63. Haskell JA. Adjusting adolescents' vocal self-perception. *Lang Speech Hear Serv Schools.* 1991;22:168-172.
64. Baken RJ, Orlikoff RF. Voice measurement: Is more better? *Logoped Phoniatr Vocol.* 1997;22:147-151.