DEFINING AND MEASURING VOICE QUALITY

Jody Kreiman[1], Diana Vanlancker-Sidtis[2], & Bruce Gerratt[1]

[1]Division of Head and Neck Surgery, School of Medicine,
University of California, Los Angeles, Los Angeles, CA, USA
[2]Department of Audiology and Speech Pathology,
New York University, New York, New York, USA
jkreiman@ucla.edu

ABSTRACT

Although voices provide listeners with significant information about speakers, defining and quantifying voice quality remain elusive goals. The ANSI standard definition of quality (that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar) is often criticized because it specifies what quality is not, rather than what it is. It has also proven difficult to devise measurement protocols for quality as specified in the ANSI definition. We argue that the ANSI definition is in fact appropriate, because it treats quality as the result of perceptual processes—interactions between listeners and signals in the context of specific perceptual goals. Application of speech synthesis in method-of-adjustment tasks allows measurement of quality psychoacoustically as those aspects of the signal that allow a listener to determine that two sounds of equal pitch and loudness are different, consistent with the ANSI definition, and provides insight into the salient acoustic attributes contributing to quality. This technique holds promise for improving the reliability and validity of measures of voice quality.

WHAT IS VOICE? THE DEFINITIONAL DILEMMA

The speaking voice naturally conveys information about the speaking individual, and voice quality serves as a primary means by which speakers project their physical, psychological, and social characteristics to the world (Laver, 1980). Although voice is central to a range of human activities, it has proven difficult to provide a single, useful, all-purpose definition of voice, and several senses of the term are in common use. Voice can be defined narrowly as sound produced by vibration of the vocal folds, or broadly as essentially synonymous with speech. Details of phonation and articulation, pitch and amplitude variations, and temporal patterning all contribute to how a speaker sounds, and broad definitions of voice reflect this fact. Which stage in the voice production process receives definitional focus depends on the interest of the practitioner, experimenter, or listener.

Defining voice quality is equally problematic. The overall quality of a sound is formally defined as that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar (ANSI, 1960). By this definition, quality is multidimensional, making it difficult to operationalize the concept. Quality is also by definition a perceptual response in the particular task of determining that two sounds are dissimilar, and it is unclear how this definition might generalize to other common, seemingly-related tasks like evaluating a single stimulus. Evidence also suggests that quality may not be independent of frequency and amplitude (Melara & Marks, 1990), as the ANSI definition seemingly requires. Finally, this definition is essentially negative, stating that quality is not pitch and loudness without indicating what it does include. Such complications have led to frequent criticism of the ANSI definition, which some claim is not a definition at all.

Chagrin about this situation has led some voice researchers to adopt definitions of quality that simply echo the narrow or broad definitions of voice described above, so that voice quality is characterized in physiological terms. Consistent with narrow definitions of voice, quality may be defined as the perceptual impression created by the vibration of the vocal folds. More broadly, voice quality may be considered the perceived result of coordinated action of the respiratory system, vocal folds, tongue, jaw, lips, and soft palate. Such definitions do not specify listeners' contributions to quality, which are essential to defining what is after all a perceptual phenomenon. For example, the perceptual importance of different aspects of a voice depends on context, attention, a listener's background, and the listening task (Kreiman et al., 1992, 1993; Kreiman & Gerratt, 2001). Thus, the measured response to a given voice signal is not necessarily constant across listeners or occasions. Physiologically-based definitions cannot accommodate such effects.

The strength of the ANSI definition is that it treats sound quality as the result of a perceptual process rather than as a fixed quantity, and highlights the importance of listeners *and* signals in determining quality. Listeners usually listen to voices in order to gather information about the environment, and the information they attend to depends on their purpose and on the information available from a particular utterance. Considered in this light, the ANSI definition has distinct advantages. Voice quality may best be thought of as an interaction between a listener and a signal, such that the listener takes advantage of whatever acoustic information is available to achieve a particular perceptual goal. Which aspects of the signal are important will depend on the task, the stimulus characteristics, and the stimulus context. Further, additional variability within a given listening task may be introduced by such listener characteristics as experience, memory, and attention. Given the many kinds of information listeners extract from voice signals, it is not surprising that the features that define vocal quality vary from task to task, voice to voice, and listener to listener.

MEASURING VOCAL QUALITY

Given the difficulties inherent in defining voice and vocal quality, it is not surprising that considerable confusion also surrounds the measurement of voice quality. The psychoacoustic study of complex, multidimensional auditory signals is in its infancy (Yost et al., 1989), and little research has examined the perceptual processes listeners apply to voice signals. Research has focused instead on identifying and defining descriptive labels or features for voices. The most common approach is simply to create a long list of terms to describe voices, and then ask listeners to assess quality by indicating the extent to which a voice possesses each feature. This approach to measuring voice quality depends on descriptive traditions rather than theory, and has changed very little in nearly 2000 years. Many common terms have been in use for centuries. Familiar labels like harsh, clear, bright, smooth, weak, shrill, deep, dull, and hoarse can be found in Roman writings on oratory (Austin, 1806), as well as in modern studies of voice quality.

Redundancies and ambiguities abound in lists of terms for voice, which tend to be exhaustive rather than efficient. To address this problem, some researchers have applied factor analysis to reduce lists of overlapping features to small sets of non-redundant scales (e.g., Voiers, 1964;

Isshiki et al., 1969). This approach preserves the descriptive tradition of quality assessment, because factors are defined in terms of the underlying scales. However, well-known limitations to this approach are also apparent. First, results of factor analytic studies depend on the input scales and stimuli, so that a factor will not emerge unless that factor is represented in the set of rating scales and is also perceptually relevant for the specific voices and utterances studied. Studies often employ restricted populations of speakers, small sets of voices, and short stimuli. For example, the well-known GRBAS protocol was developed from the results of factor analyses that used only 16 speakers and 5 steady-state vowels (Isshiki et al., 1969). Idiosyncrasies in labeling the factors may also obscure differences among studies. For example, Isshiki et al. (1969) found a breathiness factor that loaded highly on the scales dry, hard, excited, pointed, cold, choked, rough, cloudy, sharp, poor, and bad, while a breathiness factor reported by Hammarberg et al. (1980) corresponded to the scales breathy, wheezing, lack of timbre, moments of aphonia, husky, and not creaky. The validity of the factors as perceptual features also depends on the validity of the underlying scales, which has never been established. Thus, even a large-scale factor analysis (or multiple analyses) will not necessarily result in a valid or reliable rating instrument for voice quality. As a result, factor analysis has not convincingly identified scales for vocal quality that are independent and valid.

Dependence on underlying descriptive terminology can be avoided by deriving perceptual features for voices through multidimensional scaling (MDS) (e.g., Gelfer, 1993; Murry & Singh, 1980). As with factor analysis, studies using MDS have produced variable results. Some of these differences can be attributed to choice of stimuli or speaker population. However, results also indicate that listeners differ both as individuals and as groups in the perceptual strategies they apply to voices (Kreiman et al., 1990, 1992), and it does not appear that any specific features are always important for characterizing the quality of all voices under all circumstances. Scaling solutions may also leave large amounts of variance unaccounted for, and published reports may explain less than half of the variance in the underlying similarity judgments, even for simple vowel stimuli (e.g., Murry & Singh, 1980). A study of pathological voice quality (Kreiman & Gerratt, 1996) suggests that this occurs because the dimensional model of quality implied by MDS and factor analysis is not a good description of how quality is perceived. In that study, MDS solutions for 80 male and 80 female speakers each accounted for less than half of the variance in the underlying data, and revealed two-dimensional solutions in which the most severely pathological voices were separated from voices with milder pathology. Analyses of the data from individual listeners accounted for more variance (56-83%). However, stimuli did not disperse in these perceptual spaces along continuous scale-like linear dimensions, but instead clustered together in groups that lacked subjective unifying percepts. Different stimuli clustered together for each listener. These results suggest that listeners lack a common notion of what constitutes similarity with respect to voice quality. If listeners lack a common perceptual space for voice quality, then a single set of perceptual features for voice quality is not likely to be discoverable.

In the absence of empirical evidence for the validity of particular descriptors or dimensions, it is unclear why some should be included, and others excluded, in a descriptive framework for vocal quality. Further, traditional descriptive labels are holistic and independent, making it difficult to understand precisely how qualities differ from one another, or how seemingly similar qualities are related. Finally, in this tradition it is often unclear how quality relates to other parts of the speech chain. In particular, there is no formal theoretical linkage between a given quality and the physiological configuration that produced it.

Vocal profile analysis (Laver, 1980) was designed in response to these limitations. In this approach, voice quality is described in terms of the global physiological configuration that (hypothetically) underlies the overall sound of a voice. Vocal profile analysis is consistent with phonetic models of speech production, and nearly exhaustive in the physiological domain. Its primary limitation is the fact that it describes voice quality in detailed terms of the supposed underlying physiological configuration, thus indicating where perceptual information about quality *might* be, without specifying which aspects are perceptually meaningful, how listeners actually use different features to assess quality, whether (or why, or when) some features might be more important than others, or how dimensions interact perceptually.

Beyond the questionable validity of dimensional and featural protocols for assessing voice, a further difficulty is their unreliability as measurement tools. Analyses of the reliability with which listeners judge individual pathologic voices indicate that listeners almost never agree in their ratings of a single voice. Even using the simplest of phonated stimuli, the likelihood that two raters would agree in their ratings of moderately pathological voices on various 7-point scales averaged 0.21 (where chance is 0.14) (Kreiman & Gerratt, 1998). Studies of rating reliability for normal voices are less common, but not more encouraging. For example, Kendall's coefficient of concordance for ratings of 20 female voices on 16 quality scales data ranged from .14 to .69 across scales, with values averaging .33 (Gelfer, 1988).

In summary, despite a long history of research, significant difficulties continue to plague traditional approaches to voice quality measurement. Such approaches suffer from possibly irresolvable issues of rating reliability and scale validity. It is not clear what (if any) features characterize quality, or how traditional descriptors, dimensions, or articulatory distinctive features relate to overall quality (broadly or narrowly construed) or to each other. Given the difficulties inherent in measuring voice quality, some authors have argued that perceptual measures of voice should be replaced with instrumental measures (Orlikoff, 1999). In contrast to perceptual measures, instrumental measures promise precision, reliability, and replicability. However, development of instrumental protocols for measuring quality ultimately depends on our ability to define quality in a way that accounts for cognitive factors that introduce measurement variability. No theory exists describing the relationships between physiology, acoustics, and vocal quality, so it is difficult to establish which instrumental measures ought to correspond to perceptually meaningful differences in vocal quality, or why such associations should exist. Existing research has been limited largely to correlational studies, which have produced highly variable results that are difficult to interpret.

ALTERNATIVES TO TRADITIONAL MEASUREMENT SYSTEMS FOR VOICE QUALITY

Finding valid and reliable alternatives to traditional voice quality scaling methods requires hypotheses about the sources of listener disagreements, so that psychophysical techniques can be applied to devise measures that reduce these disagreements. Previous studies of pathological voices suggest that traditional perceptual scaling methods are best understood as matching tasks, in which voices are compared to mental representations that serve as internal standards for the various rating scales. These idiosyncratic internal standards appear to vary both across listeners and within a given listener, with listeners' previous experience with voices and with the context in which a judgment is made. Severity of vocal pathology, difficulty isolating individual dimensions in complex perceptual contexts, task demands, and experiential factors can also influence perceptual measures of voice (Gerratt et al., 1993; Kreiman & Gerratt, 2000; Gescheider & Hughson, 1991). These factors add uncontrolled variability to scalar ratings of vocal quality, and contribute to listener disagreement.

A protocol that does not rely on internal standards, and that makes it easier for listeners to focus attention appropriately and consistently, would eliminate many of these sources of listener disagreement. One such approach (Gerratt & Kreiman, 2001) applies speech synthesis in a method-of-adjustment task. This task allows listeners to vary acoustic parameters to create an auditory match to a voice stimulus. Because listeners directly compare each synthetic token they create to the target voice, they need not refer to mutable internal standards for particular voice qualities. Further, listeners manipulate acoustic parameters directly and hear the result of their manipulations immediately, which helps listeners focus their attention consistently. In theory, this method should improve agreement among listeners in their assessments of voice quality because it controls variance in quality judgments.

This method of quality measurement also provides other practical advantages. First, the relationship between acoustic parameters and what a listener hears is established directly, rather than correlationally. Thus, measuring quality with synthesis can experimentally establish the perceptual validity of different acoustic measures of voice. Finally, this approach to quality measurement follows directly from the ANSI definition of sound quality: It measures quality psychophysically as those aspects of the signal that allow a listener to determine that two sounds of equal pitch and loudness are different.

In a preliminary assessment of this method (Gerratt & Kreiman, 2001), listeners were asked to adjust the noise-to-signal ratio for pathological voices until the resulting synthetic stimuli matched the natural voices. In a separate experiment, listeners judged the noisiness of the same stimuli using a traditional 100 mm visual-analog rating scale. In the synthesizer task, only 3/120 listener responses differed from those of other listeners by more than a difference limen, for an agreement rate of 97.5%. In contrast, likelihood of agreement between two listeners in traditional noisiness ratings averaged 22%.

CONCLUSIONS

The appropriate method for measuring what listeners hear when they listen to voices remains an unresolved issue, and providing accurate, replicable, valid measures of vocal quality presents significant challenges. In our view, this problem is more likely to be resolved by developing methods that can assess the interactions between listeners and signals, rather than treating quality solely as a function of the voice signals themselves. In particular, application of classic psychophysical research methods may be of significant value in this area.

ACKNOWLEDGMENTS

This research was supported by grant DC01797 from the NIDCD.

REFERENCES

ANSI (1960) S1.1-1960, Acoustical terminology, New York: American National Standards Institute.

Austin, G., (1806) *Chironomia,* London: Cadell and Davies.

Gelfer, M.P. (1988) A multidimensional scaling study of voice quality in females, *Phonetica, 50*, 15-27.

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., Antonanzas-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,* 2560-2566.

Gescheider, G.A. & Hughson, B.A. (1991) Stimulus context and absolute magnitude estimation: A study of individual differences, *Perception and Psychophysics*, *50*, 45-57.

Hammarberg, B., Fritzell, B., Gauffin, J., Sundberg, J. & Wedin, L. (1980) Perceptual and acoustic correlates of abnormal voice qualities, *Acta Otolaryngologica (Stockholm), 90*, 441-451.

Isshiki, N., Okamura, H., Tanabe, M. & Morimoto, M. (1969) Differential diagnosis of hoarseness, *Folia Phoniatrica*, *21*, 9-19.

Kreiman, J. & Gerratt, B.R. (1996) The perceptual structure of pathologic voice quality, *Journal of the Acoustical Society of America, 100*, 1787-1795.

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

Kreiman, J. & Gerratt, B.R. (2000) Sources of listener disagreement in voice quality assessment, *Journal of the Acoustical Society of America, 108*, 1867 – 1879.

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

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

Laver, J. (1980) *The Phonetic Description of Voice Quality,* Cambridge, Cambridge University Press.

Melara, R.D. & Marks, L.E. (1990) Interaction among auditory dimensions: Timbre, pitch, and loudness, *Perception and Psychophysics, 48*,169-178.

Murry, T. & Singh, S. (1980) Multidimensional analysis of male and female voices, *Journal of the Acoustical Society of America, 68*, 1294-1300.

Orlikoff, R. (1999) The perceived role of voice perception in clinical practice, *Phonoscope, 2*, 87-106, 1999.

Voiers, W.D. (1964) Perceptual bases of speaker identity, *Journal of the Acoustical Society of America, 36*, 1065-107.

Yost, W., Braida, L., Hartmann, W., Kidd, G. Jr., Kruskal, J., Pastore, R., Sachs, M., Sorkin, R., & Warren, R. (1989) *Classification of Complex Nonspeech Sounds,* Washington, D.C.: National Academy Press.