VOQUAL'03, Geneva, August 27-29, 2003

Defining and Measuring Voice Quality

Jody Kreiman, Diana Vanlancker-Sidtis**, and Bruce R. Gerratt**

*Division of Head and Neck Surgery, School of Medicine
UCLA, Los Angeles
and
**Department of Audiology and Speech Pathology
New York University, New York
jkreiman@ucla.edu

Abstract

Although voices provide listeners with significant information about speakers, defining and measuring voice quality remain elusive goals. We argue that the much-maligned ANSI standard definition of sound quality is in fact an appropriate definition, because it treats quality as the result of a perceptual process rather than a fixed quantity, and highlights the interaction between listeners and signals in determining quality in the context of specific perceptual goals. Which aspects of the signal are important will depend on the task, the characteristics of the stimuli, the listener's background, perceptual habits, and so on. Given the many kinds of information listeners extract from voice signals, it is not surprising that these characteristics vary from task to task, voice to voice, and listener to listener. 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, and holds promise for improving the reliability and validity of measures of voice quality.

1. What is voice? The definitional dilemma

The speaking voice naturally conveys information about the speaking individual. The impressions listeners gain from voices are not necessarily accurate, but nevertheless voice quality serves as a primary means by which speakers project their identity—their "physical, psychological, and social characteristics"—to the world [1]. The measurement of vocal quality thus plays an important role in many disciplines, and topics related to the perception and measurement of voice quality have implications for fields ranging from evolutionary biology to music to law enforcement to medicine. Such topics encompass much of human existence, and indicate how central voice quality is to us.

It has proven difficult to provide a single, useful, all-purpose definition of voice, in part because of the broad range of functions voice subserves. As Sundberg noted [2], everyone knows what voice is until they try to pin it down, and several senses of the term are in common use. Definitions of voice fall into two general classes. Voice can be narrowly defined as "sound produced by vibration of the vocal folds," excluding the effects of vocal tract resonances, vocal tract excitation from turbulent noise, and everything else that occurs during speech production. Because anatomical constraints make it difficult to study voice as narrowly defined, most authors adopt the practical expedient of controlling for the effects of the vocal tract on voice by restricting voice samples to steady state vowels (usually /a/). This practice allows experimenters to study natural-sounding phonation, while holding non-laryngeal factors constant. This approach is the most common implementation of narrow definitions of voice.

Voice can also be broadly defined as essentially synonymous with speech. Besides details of phonatory quality, factors such as articulatory details, pitch and amplitude variations, and temporal patterning all contribute to how a speaker "sounds". Broad definitions of voice reflect this fact, and generally portray voice as the result of a complex sequence of cognitive, physiological, aerodynamic, and acoustic events. The information that complete voice patterns convey (more or less successfully) about affect, attitude, psychological state, pragmatics, grammatical function, sociological status, and personal identity emerges from this complex enfolding of phonatory, phonetic, and temporal detail. Precisely which stage in this chain of events receives definitional focus depends on the interest of the practitioner or experimenter, or on the task faced by the listener. For example, surgeons typically approach voice in terms of physiological function, with secondary concern for the exact perceived quality that results from phonation. Engineers are often interested in the acoustic waveform that correlates with vocal sound, and therefore define voice in terms of acoustic attributes. In contrast, psychologists are not especially interested in how the voice is physically produced, but instead define voice in terms of what a listener hears.

Defining voice quality is as problematic as defining voice. The overall quality (or timbre) of a sound is traditionally 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" [4]. By this definition, quality is multidimensional, including the spectral envelope and its changes in time, fluctuations of amplitude and fundamental frequency, and the extent to which the signal is periodic or aperiodic [5]. This large number of degrees of freedom makes it difficult to operationalize the concept of quality, particularly across tasks. According to the ANSI definition, quality is 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 recognizing a speaker or evaluating a single stimulus. Evidence [6] also suggests that quality may not be independent of frequency and amplitude, as the ANSI definition seemingly requires. Finally, this definition is essentially negative: It states that quality is not pitch and loudness, but does not indicate what it does include [5]. Such complications have led to frequent criticism of the ANSI definition, which some claim amounts to no definition at all [7].

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 very little to 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 other factors [8, 9], and is affected by the listening task [9, 10]. Thus, the measured response to a given voice signal is not necessarily constant across listeners or occasions.

Some of the difficulty that arises when contemplating the nature of quality may be due to the fact that quality is often treated as analogous to pitch and loudness. Authors often discuss *the* pitch or *the* loudness of a signal, presumably because these factors can be scaled unidimensionally, from low to high or faint to strong. However, quality is multidimensional, and does not possess a unique acoustic determinant. Given this fact, the perceptual effect created by a voice signal will always depend on factors like task demands, and listener attention will vary across the multiple facets of the signal, so that some are more important than others from occasion to occasion. For this reason, a single perceived quality does not necessarily consistently result from a given signal, relative to the listener. In contrast, pitch and loudness do not ordinarily vary significantly in this way, because of their unidimensional nature.

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 both 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 actually available from a particular utterance. Considered in this light, the ANSI definition has distinct advantages; in fact, its limitations can be reduced by broadening the definition to include different tasks, rather than narrowing its focus to include only a small set of specific acoustic variables. 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 characteristics of the stimuli, the listener's background, perceptual habits, and so on. Given the many kinds of information listeners extract from voice signals, it is not surprising that these characteristics vary from task to task, voice to voice, and listener to listener.

Studies of familiar voice recognition [11] highlight the importance of signal/listener interactions in voice perception. Specific articulatory information is key to identifying some individual voices, but not relevant to others [12]. Perceptual processing of voice quality differs qualitatively depending on whether the listener is familiar or unfamiliar with the voice [13]. Listeners' perceptual strategies can thus be expected to vary depending on the differential familiarity of the voices. Listeners' attention to different cues to voice identity also depends on the total voice pattern in which the cue operates [12, 14], so that the importance of a single cue varies across voices as well as listeners. Definitions of quality that focus on aspects of production or on the signal cannot account for such effects. Voice quality is the product of perceptual processes, and must be defined in terms of both signals and listeners.

2. Measuring vocal quality

Given the difficulties inherent in defining voice and vocal quality, it is not surprising that considerable confusion also surrounds the problem of measuring voice quality. The psychoacoustic study of complex, multidimensional auditory signals is in its infancy [6, 15], and little research has examined the perceptual processes listeners apply to voice signals. Research has focused instead on identifying and defining descriptive labels or distinctive features systems for voices. In this approach, vocal quality is treated as if it can be decomposed into a set of specific qualities, whose presence or absence characterizes a speaker's voice.

The most common approach to the problem of specifying voice quality is simply to create a long list of terms to describe listeners' impressions. Listeners then assess quality by indicating the extent to which a voice possesses each feature. Terms in such lists tend to be rather mixed in their level of description, and may describe voices visually (e.g., brilliant, dark), kinesthetically (strained, tight), physically (heavy, thin, pointed), aesthetically (pleasing, faulty), with reference to anatomy (pectoral, nasal), and so on [16].

This approach to measuring voice quality depends on descriptive traditions rather than theory, and has changed very little in nearly 2000 years. Familiar terms like harsh, clear, bright, smooth, weak, shrill, deep, dull, thin, hoarse, and metallic can be found in Roman writings on oratory [17; cited by 18], and also in modern studies of voice quality [19]. Some differences do exist between venerable and modern descriptive terminology. For example, antique descriptive schemes [17; cited by 18] included terms related to the personality and emotional state of the speaker (confused, doleful) and terms related to articulation and rhetorical ability (articulate, distinct). More modern compendia [19] include terms like "breathy" and "nasal" that are commonly used in the study of vocal pathology. However, similarities among traditions far outweigh differences, and many labels have been in consistent use for centuries.

Redundancies and ambiguities abound in such lists of terms, which tend to be exhaustive rather than efficient. To address this problem, some researchers have applied factor analysis to reduce large lists of overlapping features to small sets of non-redundant scales [20, 21, 39, 43]. Voice feature schemes derived from factor analysis do have obvious advantages over large lists of terms. Such analyses typically produce only a few factors, and thus measurement protocols based on such analyses [39, 43] are manageable for listeners and investigators alike. In theory, factors are independent of one another, reducing concerns about redundancies or overlap across scales, while at the same time they capture much of the information in the scalar ratings, so economy is achieved with minimal loss of information. Finally, this approach preserves the descriptive tradition of quality assessment, because factors are defined in terms of the underlying scales. Thus, factor analytic approaches bring the impression of scientific rigor to the familiar descriptive approach to quality assessment.

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 [39, 40]. These restrictions significantly limit the extent to which results can be generalized to the full spectrum of vocal qualities. Idiosyncrasies in labeling the factors may also obscure differences among studies. For example, in studies of pathological voice quality Isshiki et al. [39] 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. [43] 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. Finally, perceptual factors have been reported that are related to reliable constant listener biases and to interactions between specific voices and listeners [20]. Emergence of such factors suggests that an adequate perceptual model cannot be framed solely in terms of the stimuli, but must also account separately for differences among listeners. Overall, it thus appears that 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) [22-25]. In MDS, a perceptual space for voices is derived from listeners' ratings of how similar voices are to one another (more similar = closer together in the space). By examining correspondences between the dimensions of the derived space and measured or rated characteristics of the voices, exploratory MDS can reveal how overall vocal quality (as it determines similarities between voices) relates to scales for particular qualities. Discovery of a dimension that is highly associated with some specific quality provides evidence for the "psychological reality" of that particular quality as an important vocal feature.

As with factor analytic studies, applications of MDS to normal vocal quality have produced variable results across studies. Some of these differences can be attributed to choice of stimuli (vowel vs. sentence) or speaker population (e.g., male vs. female). However, results also indicate that listeners differ both as individuals and as groups in the perceptual strategies they apply to voices [8, 24, 26], 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 [25, 27]. This may occur because of the limited resolution of MDS: The number of extractable dimensions depends on the number of stimuli studied, which generally equals twenty or less. Alternatively, large amounts of variance may remain unexplained because the dimensional model of quality implied by MDS and factor analysis is not a good description of how quality is perceived.

A study of pathological voice quality [28] supports the latter explanation. In that study, listeners judged the similarity of all possible pairs of vowel productions obtained from sets of 80 male and 80 female speakers. Scaling solutions for male and 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 voices clustered together for each listener, and no two voices ever occurred in the same cluster for all listeners. This suggests that listeners lacked a common notion of what constitutes similarity with respect to voice quality, even when stimuli are restricted to vowels so that quality is narrowly defined. If listeners lack a common perceptual space for voice quality in its most restricted sense, then a single set of perceptual features for voice quality more broadly defined 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, rather than forming a decomposed, permutable set. This makes 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 (although terms like "nasal" may imply that such a linkage exists).

Voice profile analysis, the system of phonetic/articulatory features for voice quality proposed by Laver [1, 29, 30], was designed in response to these limitations. In this approach, voice quality is characterized as quasi-permanent and derived cumulatively throughout an individual's vocal sound production. It is then described in terms of the global physiological configuration that (hypothetically) underlies the overall sound of a speaker's voice. Laryngeal and supralaryngeal aspects of voice are both specified, and are assumed to be auditorily separable. Vocal profile analysis is analytic, consistent with phonetic models of speech production, and nearly exhaustive in the physiological domain. Because quasi-independent features (or "settings") can combine in different ways, the system can be used to describe a broad range of voice qualities in a single framework, rather than applying vague terms whose relationships to each other are unclear. Thus, for example, "hoarse" voice might appear in this system as "deep, (loud), harsh/ventricular, whispery voice," or "gruff" voice might become "deep, harsh, whispery, creaky voice" [31]. The primary limitation of this system is the fact that it models perception in terms of speech production processes without established or documented reference to a listener. That is, by describing voice quality in detailed terms of the supposed underlying physiological configuration, profile analysis indicates where perceptual information about quality *might* be. However, it does not specify which of the many aspects specified are meaningful, or, indeed, perceptible to listeners, 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. The assumption that listeners are able to separate different features auditorily is also questionable, particularly given recent evidence that listeners have difficulty isolating individual dimensions of complex voice patterns [32].

3. Limitations of traditional quality assessment protocols

The results reviewed above indicate that the validity of dimensional and featural protocols for assessing voice quality remains questionable. These protocols model voice quality solely in terms of the voice itself, although couching many of the descriptive labels in perceptual terms. They also assume an ideal and fixed listener. Most of these approaches imply that voice quality can reasonably be represented as a list or grouping of descriptors or dimensions—that there is a list of attributes that listeners can and do attend to, and that the same set describes all voices. Whether quality is broadly or narrowly construed, such frameworks imply that there exists a well-defined perceptual space for voice quality, applicable to all voices and true for all listeners, which listeners all exploit in essentially the same way.

However, substantial evidence and theoretical considerations contradict these requirements. A well-defined, theoretically motivated set of features for voice has not emerged, despite many years' research; and listeners apparently exploit vocal signals in varying ways. Data thus suggest that a fixed, common set of perceptual dimensions for voices cannot be defined, so efforts to specify a perceptually valid set of scales for voice quality are unlikely to succeed.

A further difficulty with dimensional protocols 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) [33]. The voice profile analysis system is also less than perfectly reliable, with only 52% - 65% of items in a post-test rated within one scale value of a target score after three days' training [34]. 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 [19].

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 validity. It is not clear what (if any) features characterize quality, or how traditional descriptors or dimensions relate to overall quality (broadly or narrowly construed) or to each other. Articulatory distinctive-feature approaches are analytical and motivated by phonetic theory, but they enumerate articulatory possibilities without accommodating listeners' behavior. Featural systems in general suffer from this limitation, because they model quality as if it inheres in voices, without also accounting for such listener-dependent factors as attention, experience, and response bias.

Given the difficulties, both theoretical and operational, inherent in measuring voice quality, some authors (particularly those studying pathological voices) have argued that perceptual measures of voice should be replaced with instrumental measures [16]. In contrast to perceptual measures, instrumental measures of acoustic, aerodynamic, or physiological events promise precision, reliability, and replicability. Considerations like these have motivated several measurement systems for voice, including the Dysphonia Severity Index [41] and the Hoarseness Diagram [42]. 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. Although it might be possible to define objective methods to quantify specific quality dimensions, it is more difficult to set up general rules specifying which dimensions are selected and how they combine to produce a final evaluative judgment [35]. Further, 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.

4. Alternatives to dimensional and featural 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 control such variability. Previous studies of pathological voices [9, 32] 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 with listeners' previous experience with voices [26, 36] and with the context in which a judgment is made [9, 37], and vary substantially across listeners as well as within a given listener [9, 38]. Severity of vocal pathology, difficulty isolating individual dimensions in complex perceptual contexts, task demands, and experiential factors can also influence perceptual measures of voice [3, 32]. 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 [10] 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 internal standards for particular voice qualities. Further, listeners can manipulate acoustic parameters and hear the result of their manipulations immediately. Such manipulations bring the particular acoustic dimension to the foreground, helping listeners focus their attention consistently. In theory, this method should improve agreement among listeners in their assessments of voice quality relative to traditional rating scale techniques, 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. Mappings between acoustics and quality also mean that hypotheses can be tested about the perceptual relationships between different signals. The perceptual importance of different parameters can also be evaluated in naturally occurring complex multivariate contexts. Finally, this approach to quality measurement follows directly from the ANSI definition of sound quality, in that 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 this method, listeners also create a direct mapping between the acoustic signal and a perceptual response, thus modeling quality as a process, not as a fixed entity. These characteristics suggest that the method should provide measures of quality that are valid as well as reliable.

In a preliminary assessment of this method [10], listeners were asked to adjust the noise-to-signal ratio for 12 pathological voices so that the resulting synthetic stimuli matched the natural voices as closely as possible. In a separate experiment, listeners judged the noisiness of the same stimuli using a traditional 100 mm visual-analog rating scale whose two ends were labeled "no noise" and "extremely noisy." 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%.

5. 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. Although voice quality is a psychoacoustic phenomenon, understanding of voice quality has not received benefit of classic psychophysical research methods. Pitch and loudness can often be treated as if they were functions of the signal, because measures of frequency and intensity are fairly well correlated with listeners' perceptual judgments. However, this simplification is inappropriate in the case of quality, because quality is multidimensional and listeners are flexible and variable. This is the case even when the definition of voice is constrained to refer only to laryngeal aspects of sound production. The complexities multiply with broader definitions of voice and voice quality.

Issues of quality measurement have implications beyond the study of quality itself. Once the relationship between a signal and a percept is understood, it may be possible to determine which physiological parameters create perceptually meaningful changes in phonation. At present, it is not possible to determine which aspects of vocal physiology are perceptually important, in part because the relationship between perception and acoustics (which links production to perception in the "speech chain") is poorly understood. Correlations between acoustic measures of voice and many kinds of listener judgments remain hard to interpret. Better methods of quality assessment have important implications for understanding aspects of normal voice perception (age, gender, identity, etc.) that are based in physiology, extending them to the impact of habitual speech patterns on listeners' perceptions. An improved understanding of the issues surrounding measurement of vocal quality is a first step toward these broader goals.

6. Acknowledgments

This research was supported in part by grant DC01797 from the National Institute on Deafness and Other Communicative Disorders. An greatly expanded version will appear in *The Handbook of Speech Perception*, edited by D.B. Pisoni and R. E. Remez (Blackwell, London).

7. References

[1] Laver, J., *The Phonetic Description of Voice Quality,* Cambridge University Press, Cambridge, 1980.

[2] Sundberg, J., *The Science of the Singing Voice,* Northern Illinois University Press, DeKalb, IL, 1987.

[3] de Krom, G., "Consistency and Reliability of Voice Quality Ratings for Different Types of Speech Fragments", *J.Speech Hear. Res.,* 37:985-1000, 1994.

[4] ANSI, S1.1-1960, "Acoustical terminology", American National Standards Institute, New York, 1960.

[5] Plomp, R., *Aspects of Tone Sensation,* Academic Press, London, 1976.

[6] Melara, R.D., and Marks, L.E., "Interaction Among Auditory Dimensions: Timbre, Pitch, and Loudness", *Percep. Psychophys.,* 48:169-178, 1990.

[7] Bregman, A.S., *Auditory Scene Analysis,* MIT Press, Cambridge, MA, 1990.

[8] Kreiman, J., Gerratt, B.R., Precoda, K., and Berke, G.S., "Individual Differences in Voice Quality Perception", *J.Speech Hear.Res.,* 35:512-520, 1992.

[9] Gerratt, B.R., Kreiman, J., Antonanzas-Barroso, N., and Berke, G.S., "Comparing Internal and External Standards in Voice Quality Judgments", *J.Speech Hear.Res.,* 36:14-20, 1993.

[10] Gerratt, B.R., and Kreiman, J., "Measuring Vocal Quality With Speech Synthesis", *J.Acoust.Soc.Am.,* 110:2560-2566, 2001.

[11] Van Dommelen, W.A., "Acoustic Parameters in Human Speaker Recognition",. *Lang. Speech*, 33:259-272, 1990.

[12] Van Lancker, D., Kreiman, J. and Wickens, T., "Familiar Voice Recognition: Parameters and Patterns. Part II: Recognition of Rate-Altered Voices", *J.Phonetics*, 13:39-52, 1985.

[13] Kreiman, J. "Listening to Voices: Theory and Practice in Voice Perception Research." In *Talker Variability in Speech Processing*, ed. by K. Johnson and J. Mullennix, Academic Press, New York, pp. 85–108, 1997.

[14] Van Lancker, D., Kreiman, J. and Emmorey, K., "Familiar Voice Recognition: Patterns and Parameters. Part I. Recognition of Backwards Voices", *J.Phonetics*, 13:19-38, 1985.

[15] Yost, W., Braida, L., Hartmann, W., Kidd, G. Jr., Kruskal, J., Pastore, R., Sachs, M., Sorkin, R., and Warren, R. *Classification of Complex Nonspeech Sounds,* National Academy Press, Washington, D.C., 1989.

[16] Orlikoff, R., "The Perceived Role of Voice Perception in Clinical Practice", *Phonoscope,* 2:87-106, 1999.

[17] Pollux, J., *Onomasticon.* Amsterdam edition, 1706.

[18] Austin, G., *Chironomia,* Cadell and Davies, London, 1806. Reprinted by Southern Illinois University Press, Carbondale, IL, 1966.

[19] Gelfer, M.P., "Perceptual Attributes of Voice: Development and Use of Rating Scales", *J.Voice,* 2:320-326, 1988.

[20] Voiers, W.D., "Perceptual Bases of Speaker Identity", *J.Acoust.Soc.Am.,* 36:1065-1073, 1964.

[21] Fagel, W.P.F., van Herpt, L.W.A., and Boves, L., "Analysis of the Perceptual Qualities of Dutch Speakers' Voice and Pronunciation", *Speech Communication,* 2:315-326, 1983.

[22] Matsumoto, H., Hiki, S., Sone, T., and Nimura, T., "Multidimensional Representation of Personal Quality of Vowels and Its Acoustical Correlates", *IEEE Trans. Audio Electroacous.,* AU-21:428-436, 1973.

[23] Kreiman, J., and Papcun, G., "Comparing Discrimination and Recognition of Unfamiliar Voices", *Speech Communication,* 10:265-275, 1991.

[24] Gelfer, M.P., "A Multidimensional Scaling Study of Voice Quality in Females", *Phonetica,* 50:15-27, 1993.

[25] Murry, T., and Singh, S., "Multidimensional Analysis of Male and Female Voices", *J.Acoust.Soc.Am.,* 68:1294-1300, 1980.

[26] Kreiman, J., Gerratt, B.R., and Precoda, K., "Listener Experience and Perception of Voice Quality", *J.Speech Hear.Res.,* 33:103-115, 1990.

[27] Murry, T., Singh, S., and Sargent, M., "Multidimensional Classification of Abnormal Voice Qualities", *J.Acoust.Soc.Am.,* 61:1630-1635, 1977.

[28] Kreiman, J., and Gerratt, B.R., "The Perceptual Structure of Pathologic Voice Quality", *J.Acoust.Soc.Am.,* 100:1787-1795, 1996.

[29] Laver, J., "Phonetic Evaluation of Voice Quality", in *Voice Quality Measurement*, edited by R.D. Kent and M.J. Ball, Singular, San Diego, p. 37-48, 2000.

[30] Ball, M. J., Esling, J., and Dickson, C., "The Transcription of Voice Quality", In *Voice Quality Measurement*, edited by R. D. Kent and M. J. Ball, Singular, San Diego, p. 49-58, 2000.

[31] Laver, J., "Voice Quality and Indexical Information", *British Journal of Disorders of Communication,* 3:43-54.

[32] Kreiman, J., and Gerratt, B.R., "Sources of Listener Disagreement in Voice Quality Assessment", *J.Acoust.Soc.Am.,* 108:1867 – 1879, 2000.

[33] Kreiman, J., and Gerratt, B. R., "Validity of Rating Scale Measures of Voice Quality", *J.Acoust.Soc.Am.,* 104:1598-1608, 1998.

[34] Wirz, S., and Mackenzie Beck, J., "Assessment of Voice Quality: The Vocal Profiles Analysis Scheme", in *Perceptual Approaches to Communication Disorders*, edited by S. Wirz, Whurr, London, pp. 39-55, 1995.

[35] Bodden, M., "Instrumentation for Sound Quality Evaluation", *Acustica*, 83:775-783, 1997.

[36] Verdonck-de Leeuw, I. M., "Perceptual Analysis of Voice Quality: Trained and Naive Raters, and Self-Ratings", in *Proceedings of Voicedata98 Symposium on Databases in Voice Quality Research and Education*, edited by G. de Krom, Utrecht Institute of Linguistics, Utrecht, pp. 12-15, 1998.

[37] Gescheider, G.A., and Hughson, B.A., "Stimulus Context and Absolute Magnitude Estimation: A Study of Individual Differences", *Perception and Psychophysics*, 50:45-57, 1991.

[38] Kreiman, J., Gerratt, B.R., Kempster, G.B., Erman, A., and Berke, G.S., "Perceptual Evaluation of Voice Quality: Review, Tutorial, and a Framework for Future Research", *J.Speech Hear.Res.,* 36:21-40, 1993.

[39] Isshiki, N., Okamura, H., Tanabe, M., and Morimoto, M., "Differential Diagnosis of Hoarseness," *Folia Phoniatrica*, 21:9-19, 1969.

[40] Hirano, M. *Clinical Examination of Voice*, Springer, New York, 1981.

[41] Wuyts, F., deBodt,M., Molenberghs, G., Remacle, M., Heylen, L., Millet, B., vanLierde, K., Raes, J., and vandeHeyning, P., "The Dysphonia Severity Index: An Ojective Measure of Vocal Quality Based on a Multiparameter Approach", *J.SpeechLang.Hear.Res.*, 43:796-809, 2000.

[42] Frohlich, M., Michaelis, D., Strube, H., and Kruse, E., "Acoustic Voice Analysis by Means of the Hoarseness Diagram", *J.SpeechLang.Hear.Res.*, 43:706-720, 2000.

[43] Hammarberg, B., Fritzell, B., Gauffin, J., Sundberg, J., and Wedin, L., "Perceptual and Acoustic Correlates of Abnormal Voice Qualities", *Acta Otolaryngol. (Stockholm),* 90:441-451, 1980.