

Sources of listener disagreement in voice quality assessment

Jody Kreiman^{a)} and Bruce R. Gerratt

Division of Head and Neck Surgery, UCLA School of Medicine, 31-24 Rehabilitation Center, Los Angeles, California 90095-1794

(Received 10 April 2000; accepted for publication 23 June 2000)

Traditional interval or ordinal rating scale protocols appear to be poorly suited to measuring vocal quality. To investigate why this might be so, listeners were asked to classify pathological voices as having or not having different voice qualities. It was reasoned that this simple task would allow listeners to focus on the kind of quality a voice had, rather than how much of a quality it possessed, and thus might provide evidence for the validity of traditional vocal qualities. In experiment 1, listeners judged whether natural pathological voice samples were or were not primarily breathy and rough. Listener agreement in both tasks was above chance, but listeners agreed poorly that individual voices belonged in particular perceptual classes. To determine whether these results reflect listeners' difficulty agreeing about single perceptual attributes of complex stimuli, listeners in experiment 2 classified natural pathological voices and synthetic stimuli (varying in f_0 only) as low pitched or not low pitched. If disagreements derive from difficulties dividing an auditory continuum consistently, then patterns of agreement should be similar for both kinds of stimuli. In fact, listener agreement was significantly better for the synthetic stimuli than for the natural voices. Difficulty isolating single perceptual dimensions of complex stimuli thus appears to be one reason why traditional unidimensional rating protocols are unsuited to measuring pathologic voice quality. Listeners did agree that a few aphonic voices were breathy, and that a few voices with prominent vocal fry and/or interharmonics were rough. These few cases of agreement may have occurred because the acoustic characteristics of the voices in question corresponded to the limiting case of the quality being judged. Values of f_0 that generated listener agreement in experiment 2 were more extreme for natural than for synthetic stimuli, consistent with this interpretation. © 2000 Acoustical Society of America. [S0001-4966(00)01310-2]

PACS numbers: 43.71.Bp, 43.71.Gv [KRK]

I. INTRODUCTION

Rating scale measures of vocal quality are often used clinically to evaluate pathological voices, and serve as a standard of comparison for acoustic measures of voice. Traditional rating protocols use unidimensional ordinal or interval scales, and require listeners to focus selectively on specific aspects of voice (e.g., breathiness or roughness) and assess the extent to which a voice has that particular quality. Although recent evidence (Kreiman and Gerratt, 1998) indicates that such protocols may be poorly suited to measuring vocal quality, it is unclear why difficulties arise. Multidimensional scaling data (e.g., Kreiman and Gerratt, 1996) suggest that fundamental problems exist with the validity of traditional voice quality scales. In that study, similarities among voices were not well predicted by traditional rating scales, or indeed by any set of static phonetic or linguistic-style "features." Other studies (e.g., Kreiman *et al.*, 1993; Gerratt *et al.*, 1993; Kreiman and Gerratt, 1998) suggest that problems with traditional voice assessment protocols may be due to factors in addition to or instead of scale validity. For example, individual listeners are reasonably self-consistent in their judgments of specific aspects of vocal quality, but across listeners more than 60% (and as much as 78%) of the variance in ratings of voices may be due to factors other than

differences among voices in the quality being rated (Kreiman and Gerratt, 1998). This discrepancy between test-retest and inter-rater reliability suggests that factors including stable long-term differences between raters in perceptual strategy, or short-term differences within and between listeners in attention to different aspects of the stimuli, may contribute to poor reliability of traditional voice rating protocols. (See Kreiman *et al.*, 1993, for a discussion of other hypothetical sources of variability in ratings of voice.)

Although existing data do not allow rater unreliability to be attributed unambiguously to any particular cause or causes, it is possible to use alternate measurement techniques to differentiate problems of scale validity from other sources of disagreements among raters. Binary classification systems for describing voice quality, in which voices are assigned to broad categories based on quality (a breathy voice; a strained voice) may offer such clues. Classification tasks differ from traditional scalar judgments in the level of measurement required and in the complexity of the judgment made, and we reasoned that these simplifications would allow listeners to focus on the *kind* of quality of a voice has, rather than how much of a quality it possesses. If traditional labels for voice quality have psychological reality, then listeners should agree that some voices unambiguously possess that quality (a breathy voice; a rough voice).

Classification systems for describing voice quality are

^{a)} Author to whom correspondence should be addressed; electronic mail: jkreiman@ucla.edu

very old, and underlie many modern studies of voice. Such systems have their basis in studies of oratory and elocution, with many common terms dating from the Romans. For example, in the first century BC Cicero (see Cicero, 1948) used the term “asperam” (“rough”), and in the second century AD Julius Pollux (see Pollux, 1706; cited by Austin, 1806) employed terms like “aeneam” (“brassy”) and “raucam” (“hoarse”) (Austin, 1806; Laver, 1980). Attempts at systematic classification of vocal quality began in the 19th century. For example, Rush (1859) distinguished four qualities of voice (“natural,” “falsette,” “whispering,” and an “improved” quality labeled “orotund”), and also described “guttural vibration” and “tremulous movement.” Goldbury and Russell (1844; cited by Gray, 1943) distinguished the qualities harsh, smooth, aspirated, pectoral, guttural, oral, orotund, and pure tone. (For interesting reviews see Plugge, 1942; Gray, 1943; Laver, 1980).

Despite early concerns that such labels for voice are at best metaphorical (Rush, 1859), these traditional classification systems for describing voice are readily discernible in contemporary descriptive usage. Although modern classification systems for measuring vocal quality have not to our knowledge been formally proposed, common usage and many studies assume that such systems are valid. Classification of voice qualities is especially prevalent in studies of speech synthesis, because attempts at synthesizing particular qualities (e.g., Wendahl, 1966; Klatt and Klatt, 1990; Childers and Lee, 1991; Lalwani and Childers, 1991) presuppose that perceptual classes for voice exist. Authors in such studies typically sort voices into groups based on perceptual criteria, and then investigate the synthesis strategies necessary to model that kind of phonation. Thus Childers and Lee (1991) selected examples of breathy, modal, fry, and falsetto phonation, and then examined the synthesis parameters necessary to reproduce each voice type. Gobl and Ni Chasaide (1992) modeled a single normal speaker who produced modal, breathy, whispery, tense, lax, and creaky voice, and Kasuya and Ando (1991) studied the synthesis of breathiness by selecting two “breathy” voices and copying them.

Other researchers (e.g., Martin *et al.*, 1995; Hillenbrand and Houde, 1996) have used a classification step as a precursor to gathering scalar rating of vocal quality. In these studies, voices were first sorted into classes corresponding to specific pathological voice “types” (i.e., qualities), and then were rated on the extent to which they possessed that quality. For example, Hillenbrand and Houde (1996) first selected a set of voices that “appeared to depart from normal voice quality primarily in the direction of breathiness” (p. 313), after which listeners rated the level of breathiness of those stimuli.

A few researchers have examined listeners’ abilities to classify voices in the manner required by such studies. Colton and Estill (1981) asked normal speakers to produce samples of conversational speech, cry, twang, and operatic ring, which listeners then sorted into four classes with better than chance accuracy. Rammage *et al.* (1992) trained expert listeners with standard samples of several qualities (breathiness, strain/harshness, high pitch, low pitch, glottal attacks, phonation breaks, pitch breaks, and roughness/glottal fry).

They then asked listeners to rate severity of dysphonia, to check off which features were present in each voice sample, and to indicate which of the checked features was dominant in the sample. Although no statistical analysis was undertaken, listeners reportedly agreed well about the overall severity of pathology, but not about the specific perceptual features present or dominant in each sample. Martin and colleagues (Martin *et al.*, 1995; Martin and Wolfe, 1996) asked listeners to sort voices into four groups (breathy, rough, hoarse, normal) after training with synthetic prototypes for each voice type. In both studies, about 60% of listeners agreed about the class in which about 60% of voices belonged; agreement was below 60% for the remaining voices. It is not clear from their discussion if agreement exceeded chance levels.

In the present study, we examined levels and patterns of listener agreement in a binary classification task for pathological voices. In particular, we attempted to determine whether listener disagreements were due to fundamental problems of scale validity, or to the manner in which quality has been measured. If listeners agree reliably in their classifications, the validity of traditional scales is supported, and disagreement in ordinal and interval rating scale protocols may be attributed to other factors, including limitations of the particular measurement techniques employed.

II. EXPERIMENT 1

A. Method

1. Stimuli

Stimuli were drawn from a previous study (Kreiman and Gerratt, 1996), where they are described in detail. Briefly, the voices of 80 male and 80 female speakers with vocal pathology were selected from a large library of samples recorded under identical conditions as part of a phonatory function analysis. Each speaker sustained the vowel /a/ for as long as possible. Voices were recorded using a microphone placed off-axis 5 cm away from the speaker’s lips. Utterances were low-pass filtered at 8 kHz and digitized directly at 20 kHz with 12-bit resolution. A 2-s sample was excerpted from the middle of each utterance. Stimuli were equalized for peak intensity, and onsets and offsets were multiplied by 40-ms ramps to eliminate click artifacts.

Speakers ranged in age from 18–96 years, and represented a variety of diagnoses. Severity of pathology was rated on a 6-point equal-appearing interval scale by unanimous vote of the authors and an experienced speech-language pathologist. (Differences in rating were resolved by discussion.) Chi square analysis indicated that severity of pathology, diagnostic category, gender, and speakers’ ages were statistically independent in these voice sets (Kreiman and Gerratt, 1996), reducing the likelihood that differences in quality are confounded with extraneous factors.

2. Listeners

A total of 19 expert listeners participated in these experiments. Eleven listeners provided two sets of quality judgments (one for breathiness and one for roughness), and 8 listeners made judgments of one quality only, for a total of 15 listeners/task. Each listener had a minimum of 3 years’

post-graduate experience evaluating and/or treating voice disorders. Listeners reported no history of hearing difficulties.

3. Procedure

At each session, listeners heard the 160 stimulus voices, along with 40 repeated trials (inserted at random into the sequence of trials), for a total of 200 trials/session. Testing took place in a double-walled IAC sound-attenuated booth. Stimuli were low-pass filtered at 8 kHz and presented in free field over high fidelity loudspeakers (Boston Acoustics AD40) at a comfortable listening level (approximately 80 dB SPL).

Judgments of breathiness and roughness were made at separate sessions. For each voice, listeners were instructed to decide whether or not its quality departed from normal primarily in the direction of breathiness (or roughness), and to respond either “primarily breathy” or “not primarily breathy” (or “primarily rough”/“not primarily rough”). They were asked to disregard the severity of pathology or apparent type of voice disorder, and were allowed to replay each stimulus as often as necessary before making their judgment. When listeners participated in two sessions, these were separated by at least 1 week. Task order was randomized across listeners, and stimuli were rerandomized for every presentation. Each listening session lasted about 30 min.

B. Results

1. Test-retest agreement

For judgments of breathiness, test-retest agreement (the percentage of repeated voices placed in the same class both times they were presented) averaged 85.0% across listeners [standard deviation (s.d.)=6.81; range=72.5%–97.5%]. For judgments of roughness, test-retest agreement averaged 80.7% (s.d.=6.97; range=70%–92.5%).

2. Classification responses

Listeners varied in how frequently they applied the labels “primarily breathy” and “primarily rough” to the stimuli. Across listeners, the number of “rough” responses ranged from 26–101/160 (mean=63.4; s.d.=20.3). The number of “breathy” responses ranged from 33–94/160 (mean=60.8, s.d.=15.4).

Despite differences in rates of responding, pairs of listeners agreed fairly well in their classification judgments. On average, two listeners agreed in their responses for 73.5% of voices in the breathiness task (s.d.=6.3%; range=58.1%–88.1% agreement), and for 69.5% of voices in the roughness task (s.d.=7.3%; range=49.4%–100%).

However, across all listeners, agreement levels were rather poor. The frequency with which listeners responded “primarily breathy” or “primarily rough” was calculated for each voice, and the distribution of these frequencies is shown in Fig. 1. In this figure, a value of 15 on the x axis (rightmost columns) indicates that all 15 listeners agreed a voice was primarily breathy or rough; a value of 0 indicates that all listeners agreed the voice was not primarily breathy or rough (i.e., 0 listeners classified the voice as breathy or

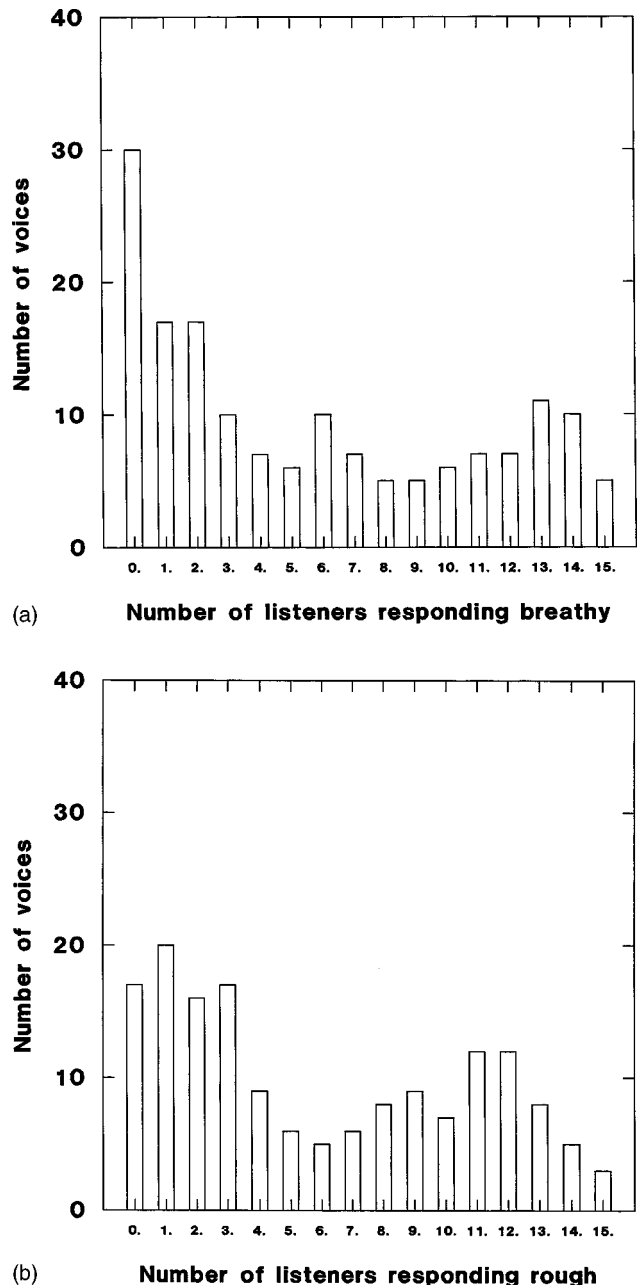


FIG. 1. Distribution of agreement levels for the two binary classification tasks. The x axis shows the number of listeners agreeing in their classification of a voice; the y axis shows the number of voices which received that level of agreement. Column totals sum to 160, the number of voice stimuli. (a) Breathiness judgments. (b) Roughness judgments.

rough). The relative height of the endpoints depends on the *a priori* distribution of voice qualities in the population, and thus cannot be interpreted directly. However, if listeners agree in their judgments that voices did or did not belong in a class, these functions should dip to zero between endpoints, because the center of the x axis represents maximum disagreement. Thus good listener agreement should result in a roughly U-shaped curve.

This was not the case in the present data. Listeners agreed better that a voice was not primarily breathy or rough than they did that a voice belonged in a class, but otherwise levels of agreement were rather flat across the scale, and did not approach zero between endpoints. Further, listeners

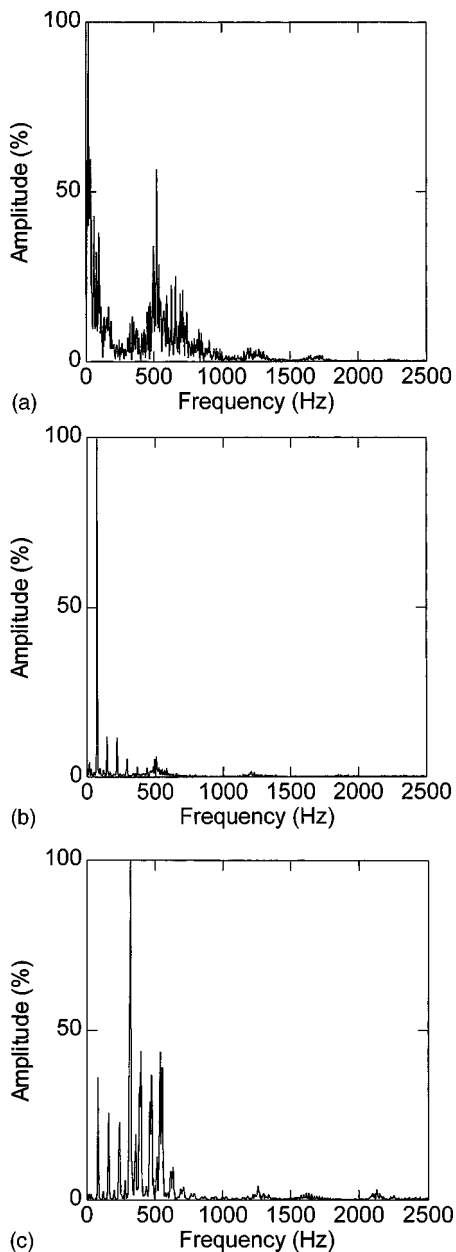


FIG. 2. Examples of linear FFT spectra for voices unanimously judged to be primarily breathy or primarily rough. Amplitude ranges from minimum to maximum, in percent. (a) A voice unanimously judged to be primarily breathy. (b) A voice unanimously judged to be primarily breathy. (c) A voice unanimously classified as primarily rough. Note the presence of interharmonics.

unanimously agreed that only 3/160 voices (1.9%) were primarily rough; only 5/160 voices (3.1%) were unanimously judged primarily breathy.

All five voices unanimously classified as “primarily breathy” were aphonic or near-aphonic, with limited harmonic structure above 700 Hz [Figs. 2(a), (b)]. The three voices unanimously classified as “primarily rough” were acoustically heterogeneous. All three were characterized by intermittent or continuous bifurcations and interharmonics; one had a very low f_0 , and one included prominent vocal fry [Fig. 2(c)]. The number of listeners responding “rough” was also significantly correlated with rated severity of vocal pathology ($r=0.70$, $p<0.05$), suggesting that roughness is also

confounded with severity of pathology. The likelihood that a voice would be judged primarily breathy is also significantly correlated with severity in these data ($r=0.49$, $p<0.05$), but the correlation is significantly lower than for roughness ($t(157)=4.15$, $p<0.05$).

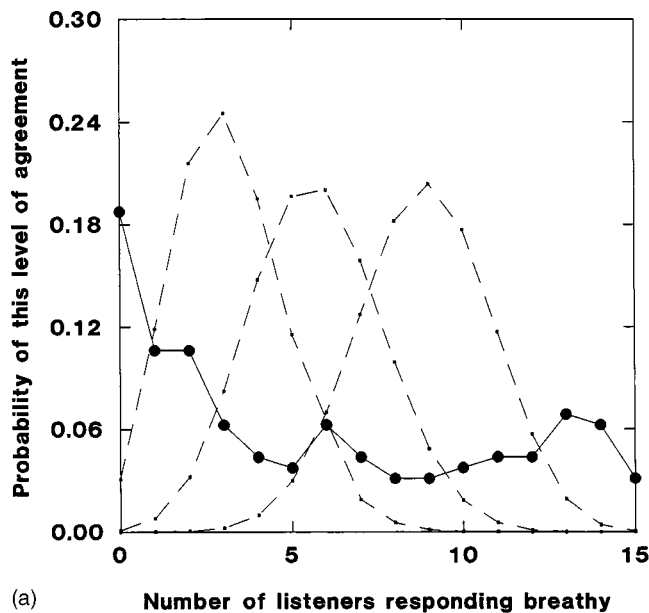
Statistical analysis of these data is complicated by the fact that there are no correct or incorrect answers. Voices cannot be placed *a priori* into perceptual classes, so the *a priori* proportions of breathy and rough voices in the population of pathological voices are unknown. For this reason, we used three different estimates of the frequency of “primarily breathy” and “primarily rough” voices in the overall population of voices to test the hypothesis that observed patterns of agreement were due to chance. The first estimate was the proportion of “primarily x ” responses by the listener who responded “primarily x ” least often; the second estimate was the proportion of responses by the listener with the largest number of “primarily x ” responses; and the third estimate was the overall proportion of “primarily breathy” or “primarily rough” responses across the pooled group of listeners. The binomial probabilities of observing N agreements, $N-1$ agreements, $N-2$ agreements, and so on, were calculated (e.g., Hays, 1994, p. 140 ff.) for each of these three different estimates of the frequency of each quality in the population of speakers.¹

Expected values (given the assumed proportion of “breathy” and “not breathy” voices in the sample and the assumption of random sampling from that population) are plotted with observed levels of agreement in Fig. 3. As this figure shows, observed agreement for both qualities is above the expected values at the margins of the figures, but below expected values in the middle of the figures. In other words, listeners agreed at above chance levels, and disagreed at below chance levels. Thus the hypothesis that results reflect random guessing can be rejected, for a range of assumed probabilities of “primarily breathy” and “primarily rough” voices.

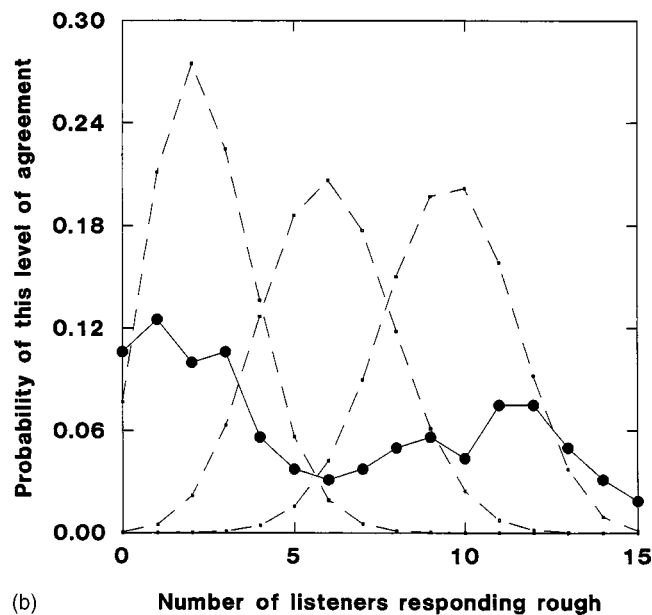
C. Discussion

These results do not support the traditional assumption that pathological voices can be meaningfully assigned to broadly applicable perceptual classes. Although listeners agreed at better than chance levels, for most voices substantial disagreement existed as to whether or not that voice belonged in a given class. However, listeners did agree in their classification judgments for a few pathological stimuli, consistent with the venerable idea that commonly used categories like breathiness and roughness are real. We speculate that listeners agreed in their judgments of these particular voices because the stimuli correspond to acoustic or physiological extremes, which represent a limit of phonation and thus are perceptually stable across listeners. When voices do not approach such a phonatory limit, we speculate, listeners are unable to consistently isolate and assess single dimensions in highly variable perceptual contexts, so perceptual strategies diverge.

For example, all five voices unanimously classified as “primarily breathy” were aphonic or near-aphonic, as noted above. Aphonia is a well-defined physiological and acoustic



(a)



(b)

FIG. 3. Expected (dashed curves) and observed (solid curves) levels of listener agreement for the binary classification tasks. The x axis indicates the number of listeners agreeing in their classification of a voice; the y axis shows the probability of that many listeners agreeing, given the assumed likelihood of “breathy” or “rough” voices in the population. The leftmost dotted curve represents the most conservative assumption about this *a priori* probability; the rightmost curve represents the most liberal assumption. (a) Judgments of breathiness. (b) Judgments of roughness.

limit to phonation—the point at which the vocal folds cease to vibrate and the vocal tract is excited solely with turbulent noise—and thus constitutes the limiting case for breathiness as classically defined (e.g., Fairbanks, 1940). Because aphonia is the extreme case of breathiness, other aspects of vocal quality (for example, fundamental frequency, overall loudness, or the spectral characteristics of the turbulence noise) are perceptually irrelevant; the voice is breathy whatever values any other characteristics may assume. Thus the listeners’ task is simplified; in essence, the voice loses degrees of freedom perceptually, so listeners lose opportunities to disagree.

Note that the values of these other aspects may be relevant when other qualities (such as the roughness of the voice) are judged, so that a single voice can be placed in more than one perceptual class. Thus phonation that is “perceptually stable” with respect to one quality is not necessarily “perceptually simple.”

In the case of roughness, vocal fry and bifurcated phonation are also well-defined physiologically and acoustically (e.g., Hollien *et al.*, 1966, 1977; Herzel *et al.*, 1991; Berry *et al.*, 1996; Omori *et al.*, 1997), and are well-distinguished perceptually from each other and from modal phonation (Hollien and Wendahl, 1968; Michel and Hollien, 1968; Omori *et al.*, 1997; see Gerratt and Kreiman, 2000, for review). Apparently these kinds of phonation are both associated with the label “rough,” which is also significantly confounded with overall severity of pathology. Variable attention by individual listeners to these different aspects of the voice signal may account for lower overall agreement that a voice was (and was not) “primarily rough.”

One concern limits our interpretation of results from this study. Although disagreements among listeners may be related to listener difficulty in agreeing about single perceptual facets of complex voice stimuli, as argued above, findings may also be artifacts of the restrictive binary classification task used. Because the task required listeners to segment continuously varying vocal quality into two discrete classes, listeners’ disagreements may reflect difficulties and differences in the placement of class boundaries. To investigate this possibility, we asked listeners to classify the natural voice stimuli used in experiment 1 according to their vocal pitch. Listeners were also asked to make similar judgments for a set of synthetic stimuli varying only in fundamental frequency. Pitch was selected as a stimulus dimension because it fulfills several criteria. First, problems regarding scale validity should not provoke disagreement among listeners, because the psychological reality of pitch is well established (see, e.g., Plomp, 1976, for review). Further, voice fundamental frequency has consistently emerged as perceptually important from studies of vocal quality (Kreiman *et al.*, 1990). Finally, the acoustic correlates of pitch are well understood, so synthesis is straightforward and accurate. If listener disagreements are related to inconsistent segmenting of a stimulus continuum, patterns and numbers of disagreements should be similar for the synthetic and natural voices, because pitch varies continuously in both cases. On the other hand, if disagreements are due to inconsistencies in how listeners isolate single dimensions in complex patterns, then agreement should be better for the synthetic stimuli than for the natural stimuli. In this view, the homogeneity of the synthetic stimuli promotes uniform perceptual strategies both within and across listeners, because only one variable changes within a fixed context across the stimulus set. Thus a consistent perceptual strategy can be adopted across listeners and applied for all the stimuli. With the natural voice samples, a single perceptual strategy is far less likely to emerge across listeners and voices, because pitch cues operate in perceptual contexts that vary widely from voice to voice. Consequently, the acoustic complexity of the pathologic voices hypothetically prevents listeners from converg-

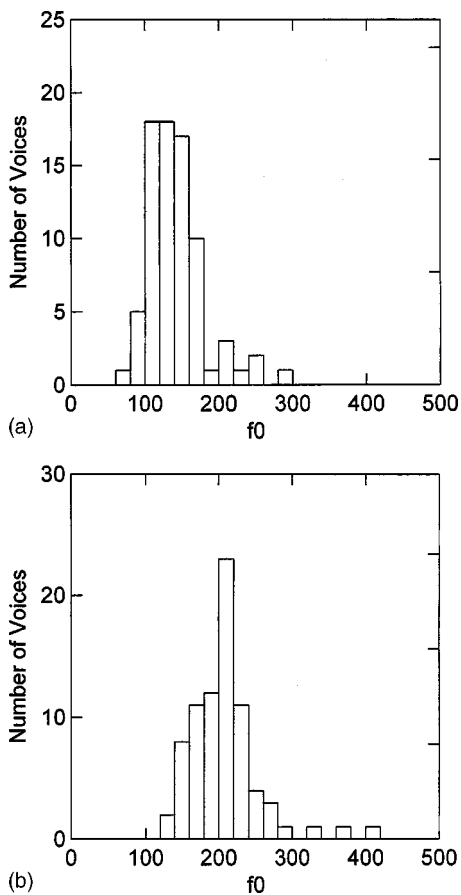


FIG. 4. Distributions of f_0 values for the natural and synthetic voice stimuli. (a) Male voices. (b) Female voices.

ing on a single perceptual strategy and making the same judgments.

III. EXPERIMENT 2

A. Method

1. Stimuli

Two sets of stimuli were used in this experiment. The first included the 80 male and 80 female natural voice samples used in experiment 1. The second included synthetic male and female voices whose f_0 values varied to match the distribution of mean f_0 values for the natural stimuli (Fig. 4). Mean f_0 values were measured from the center frequency of the lowest harmonic in a fast Fourier transform (FFT) spectrum calculated over the entire voice sample, and again with CSpeech software (Milenkovic, 1987; Milenkovic and Read, 1992). Values for voices with bifurcations or prominent amplitude modulations were rejected. The final

TABLE I. f_0 and formant frequencies for the synthetic stimuli.

	Females	Males
f_0 (Hz)	125–425	70–370
F_1 (Hz)	850	800
F_2 (Hz)	1400	1346
F_3 (Hz)	2815	2500
F_4 (Hz)	4299	3400
F_5 (Hz)	...	4373

set of synthetic voices included 77 male and 78 female tokens. Synthetic stimuli (2 s in duration) were created with a custom-designed formant synthesizer. Synthesizer parameters other f_0 were modeled on natural tokens of /a/ spoken by normal male and female speakers (Table I), and were held constant across stimuli.

2. Listeners

Fifteen expert listeners participated in this experiment. Four had participated in experiment 1; however, the two experiments were separated by several months. Each listener had a minimum of 3 years' post-graduate experience evaluating and/or treating voice disorders. Listeners reported no history of hearing difficulties.

3. Procedure

For each stimulus, listeners were asked to decide whether the voice was low pitched or not low pitched, relative to average normal speakers of the appropriate gender. They were allowed to replay stimuli as often as necessary before making their decisions. Because f_0 expectations differ for male and female voices, male and female stimuli were presented in separate blocks of trials, as were synthetic and natural stimuli. Thus each listener heard four blocks of stimuli: male natural voices, female natural voices, male synthetic voices, and female synthetic voices. Order of blocks was randomized across listeners, and stimuli within a block were rerandomized for each listener. Listeners were informed of the class of stimuli to be judged prior to hearing each block of trials.

To assess test-retest reliability, 20% of trials (selected at random) were repeated in each block of stimuli. Repeated trials were inserted at random into the sequence of trials. Other testing conditions were identical to those used in experiment 1. Listeners completed all four blocks of trials at a single session lasting about 1/2 hour.

B. Results

1. Test-retest agreement

For the natural stimuli, test-retest agreement was comparable to that observed in experiment 1, averaging 81.8% across listeners (s.d.=9.61; range=62%–95%). Listeners were significantly more self-consistent when classifying the synthetic stimuli [mean test-retest agreement=86.7%; s.d.=9.0; range=73.3%–100%; matched samples $t(14) = -2.21$, $p < 0.05$].

2. Classification responses

As with judgments of breathiness and roughness, listeners varied in the number of voices they considered low pitched. For the natural stimuli, the number of “low pitched” responses was similar to the number of “primarily breathy” and “primarily rough” responses, ranging from 40–91 (out of 160 stimuli; mean=62.5, s.d.=15.3). Pairwise agreement among raters for judgments of the pitch of natural stimuli was also similar to agreement for breathiness and roughness. On average, two listeners agreed about 73.9% of

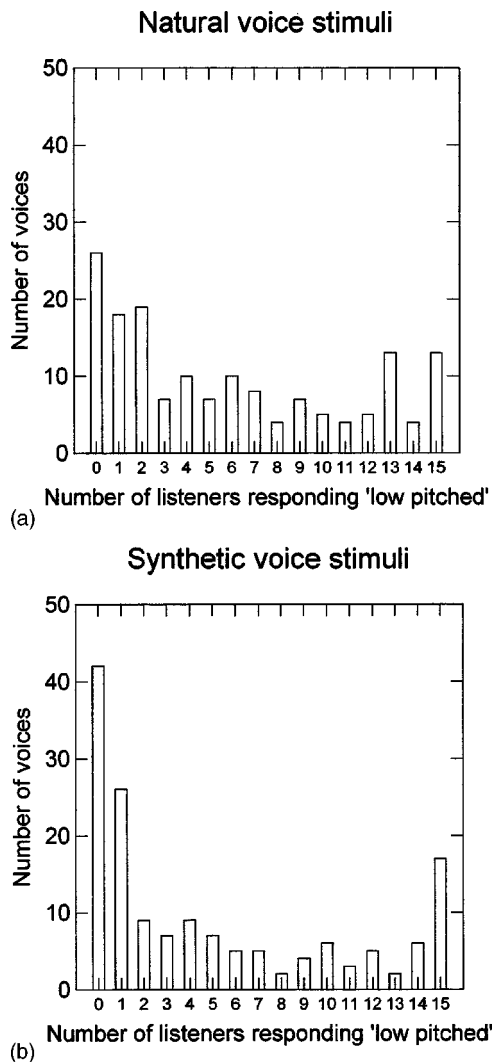


FIG. 5. Distribution of agreement levels for the pitch judgment tasks. The x axis shows the number of listeners agreeing in their classification of a voice; the y axis shows the number of voices which received that level of agreement. (a) Natural voice stimuli. Column totals sum to 160, the number of pathological voice stimuli. (b) Synthetic voice stimuli. Column totals sum to 155, the number of synthetic voice stimuli.

their pitch judgments (compared to 73.5% agreement in breathiness judgments, and 69.5% agreement in roughness judgments).

For the synthetic stimuli, the proportion of “low pitched” responses was lower than that for the natural stimuli (mean = 51.5/155 stimuli; s.d. = 18.3; range = 29–84). However, pairwise agreement among listeners was significantly better for judgments of the synthetic stimuli [mean pairwise agreement = 80.1%; s.d. = 7.6; range = 61.9%–92.3%; matched samples $t(103) = -7.85$, $p < 0.05$].

Patterns of overall agreement for the two sets of pitch judgments are shown in Fig. 5. As in experiment 1, a value of 15 on the x axis (rightmost columns) indicates that all 15 listeners agreed a voice was low pitched; a value of 0 on the x axis indicates that all listeners agreed the voice was not low pitched (i.e., 0 listeners classified the voice as low pitched).

As this figure shows, listeners agreed substantially better for the synthetic than for the natural stimuli, even though f_0 values were identical for the two sets of stimuli. Listeners

unanimously agreed that 27.1% of synthetic stimuli were not low pitched, vs 16.2% of natural stimuli; and they agreed unanimously that 11.0% synthetic stimuli were low pitched, vs 8.1% of natural stimuli.

In addition, patterns of listener agreement and disagreement differed for the two tasks. For the natural stimuli, levels of agreement were rather flat across the figure, and did not approach zero between endpoints, indicating that for many voices listeners were divided as to whether or not that voice was low pitched. In contrast, for the majority of the synthetic stimuli (91/155) all or all but one of the listeners agreed in their judgments. Significant disagreement occurred for relatively few voices, resulting in the predicted U-shaped function. Thus the observed disagreements for the synthetic stimuli seem to represent primarily minor differences in the placement of boundaries between classes of stimuli.

To test the hypothesis that agreement rates were better for judgments of synthetic stimuli than for judgments of the breathiness, roughness, or pitch of natural stimuli, data were first transformed using the following procedure. Recall that classification responses for individual voices ranged from 0 (all raters agreed the voice did not belong in a class) to 15 (all raters agreed the voice did belong in the class). This scale was modified such that it ranged from perfect agreement to maximum disagreement among raters. That is, scores of 0 and 15 were converted to 100% agreement; scores of 1 and 14 were converted to 93.3% agreement (i.e., all but one rater agreed in their classification judgment); scores of 2 and 13 were converted to 86.7% agreement; and so on. Note that this new scale ranged from 100% agreement to 53.3% agreement, because a divided panel (7 votes vs 8 votes) represented the maximum possible disagreement among raters.

Because these data can assume only a small number of values, a Kruskal-Wallis one-way nonparametric analysis of variance (ANOVA) was used to compare transformed agreement rates for the four binary classification tasks (judgments of breathiness, roughness, pitch/natural stimuli, and pitch/synthetic stimuli). Tasks differed significantly in the levels of overall agreement observed (Kruskal-Wallis test statistic = 21.21, $df = 3$, $p < 0.05$). *Post-hoc* comparisons indicated that listeners agreed significantly better ($p < 0.05$, adjusted for multiple comparisons) in their judgments of the synthetic stimuli than they did in the other three tasks, for which agreement levels did not differ ($p > 0.05$).

Figure 6 shows the likelihood of listener agreement plotted against f_0 for the natural and synthetic stimuli. For the natural voices, pitch category is apparently ambiguous for fundamental frequencies below about 300 Hz for female voices (shown as filled circles in the figure), and below about 200 Hz for male voices (shown as stars). Voices with f_0 above these values were unambiguously not low pitched; but voices with f_0 below these values might or might not be considered low pitched. A different pattern emerged for the synthetic voices. For these stimuli, the likelihood of “low pitched” responses decreased smoothly across frequencies, and bottomed out at about 150 Hz for males and 250 Hz for females.

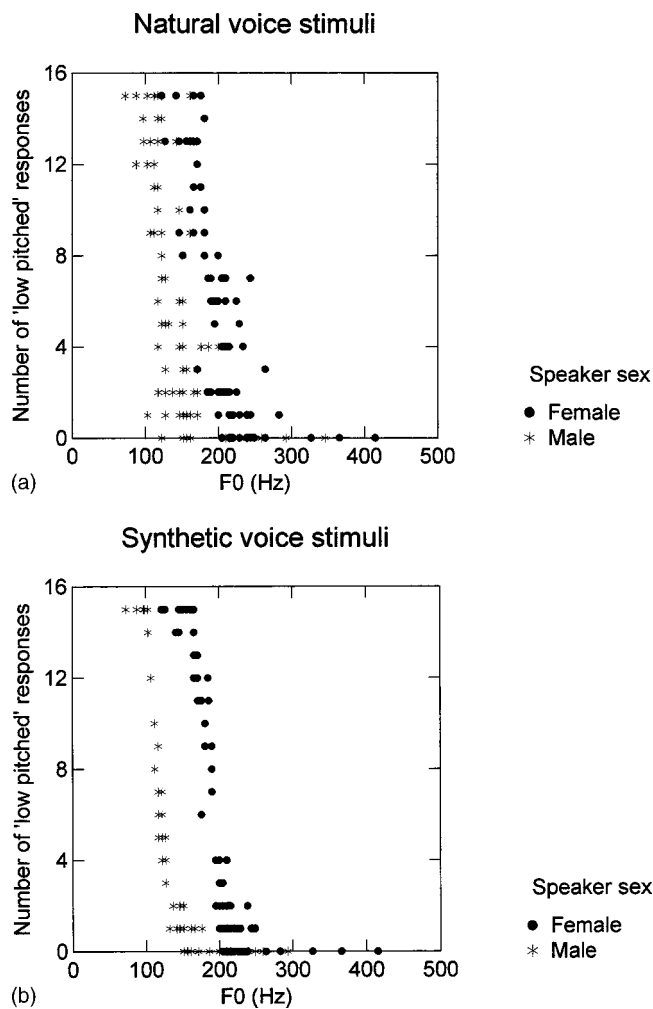


FIG. 6. The likelihood of listener agreement about the pitch of natural and synthetic stimuli, vs f_0 . Male stimuli are plotted with stars; female stimuli are plotted with filled circles. (a) Natural voice stimuli ($n=160$). (b) Synthetic voice stimuli ($n=155$). Note that many points overlap in this figure.

C. Discussion

These results indicate that the high levels of disagreement observed for judgments of natural stimuli were not due solely to difficulties segmenting a continuum or to some other task-related factor. Instead, they appear to be directly related to characteristics of the stimuli. Listener agreement about the pitch of the natural voices did not differ significantly from agreement about breathiness and roughness. However, both test-retest agreement and inter-rate agreement were significantly better when listeners classified the pitch of synthetic voice stimuli, which varied only in f_0 . Listener agreement for the synthetic voices was very well predicted by f_0 , as one would expect. Agreement increased sharply and smoothly as f_0 departed from population mean values (about 130 Hz for male speakers and 220 Hz for females; e.g., Peterson and Barney, 1952). The relationship between f_0 and pitch was more complicated for the natural voices. Although in general listener agreement varied with f_0 , listeners sometimes agreed about the pitch of voices whose f_0 was near the population mean, and disagreed about voices for which f_0 was substantially above or below average [Fig. 6(a)]. Further, the f_0 values at which listeners no longer

disagreed about vocal pitch were about 50 Hz higher for the natural than for the synthetic stimuli. Thus it appears that not only listener agreement levels, but also the amount of evidence listeners require to agree about the presence of a vocal quality, depend on stimulus complexity. These results are consistent with the hypothesis that listeners are unable to agree in their judgments of specific attributes of voice because they cannot consistently focus attention on individual quality dimensions of complex signals.

IV. GENERAL DISCUSSION

These results are consistent with the claim that unidimensional rating scale approaches are inappropriate for measuring pathological vocal quality. Even in a simple binary classification task, listeners were unable to agree with each other consistently about the breathiness, roughness, or pitch of natural pathological voice stimuli. However, agreement was significantly better for pitch judgments when stimuli were relatively simple synthetic vowels. This suggests that disagreements are not due to characteristics of the classification task, but instead are related to listeners' difficulty in isolating single dimensions of complex stimuli.

The notion that listeners agree about the quality of a few pathological voices because those voices correspond to physiological or acoustic extremes accounts well for the present data. If listeners agree in their judgments only when a voice is at or very near a phonatory limit, agreement will be uncommon, because such voices are relatively rare. This explanation also accounts for the fact that listeners agreed better about which voices were not breathy, rough, or low pitched than about which voices belonged in a particular class. As argued above, only a single, extreme, relatively uncommon vocal configuration will generate agreement that a voice belongs in a class. Agreement that a voice is not in a class requires that listeners agree that the voice does not correspond to one specific acoustic pattern, but does not require that they agree about the particular manner in which it deviates from that pattern. Such limited agreement is apparently relatively easy to achieve. Although this interpretation is speculative at present due to the small number of voices about which listeners agreed, the hypothesis that phonation near physiologic or acoustic limits is reliably perceived across listeners could be tested, using either natural or synthetic stimuli.

Because traditional dimensions for voice quality usually range from "normal" to "severe," the present results might seem to imply that valid rating protocols could be constructed with reference to these extreme stimuli. For example, such voices could hypothetically be used to create sets of "anchor stimuli" varying in steps from normal to the extreme. Listeners could then judge quality with reference to these anchors, rather than by comparison to variable internal standards for a quality (e.g., Gerratt *et al.*, 1993). However, even if a few acoustically extreme voice stimuli are treated as "cardinal" in quality (Jones, 1922; Abercrombie, 1967), these voices are unlikely to provide a basis for creating a useful set of ordinal- or interval scale features for all voices. As voices depart from the extreme limits of phonation, no basis exists for weighing the many different facets of quality

that occur, and listeners are free to focus their attention in any way they like, leading to listener disagreement. In this way, the existence of a few “cardinal” voices does not imply that continuous features for the vast majority of voices can also be defined, or that other voices can be classified or ranked with reference to these extreme stimuli.

These results also suggest that perception of pathological qualities like breathiness and roughness may differ from the perception of phonemic breathiness and roughness in languages that distinguish phonation types, analogous to differences between the perception of (nonphonemic) pitch and (phonemic) tone (e.g., Gandour and Harshman, 1978; Gandour *et al.*, 1988). Vocal quality varies continuously in many dimensions, as phonetic quality does, but no formal system of contrasts (analogous to the phonology of a language) exists to divide this voice quality continuum into discrete classes. Linguistic sound categories (phonemes) can be established by reference to contrasts in meaning, but when judging voice quality, an infinite number of arrangements is possible, so “mistake” is undefined and listeners have an unrestricted choice of responses. (See Belkin *et al.*, 1997, for discussion of a similar problem with the description of odors.)

In conclusion, the data presented here suggest that traditional labels for vocal quality may be valid in a limited way, but that pathologic voice quality assessment using traditional perceptual labels is not generally useful. The particular pattern of observed listener disagreements appears to be related in part to difficulty isolating single perceptual dimensions of complex stimuli, and listener agreements may be accounted for by the relative perceptual stability of a small number of stimuli corresponding to well-defined acoustic or physiological extremes. This pattern of agreements and disagreements among listeners is consistent with problems that have arisen in the study of other sensory modalities (for example, taste and smell) that also lack category or featural structure, and for which no satisfactory, consensually accepted descriptive terminology exists (e.g., Belkin *et al.*, 1997), and possibly cannot exist in the absence of such structure. Measuring perceptual responses to such stimuli presents a considerable challenge.

ACKNOWLEDGMENTS

We thank James Hillenbrand and Thomas Baer for insightful and enormously helpful comments on an earlier version of this paper. This research was supported by Grant No. DC01797 from the National Institute on Deafness and Other Communication Disorders.

¹With the number of listeners as N , the proportion of breathy (or rough) voices in the population as p , and q equal to $1-p$, the chance probability of N successes (N listeners responding “breathy” or “rough”), $N-1$ successes, $N-2$ successes. . . 0 successes, can be estimated using the formula $p(r \text{ successes}; N, p) = \binom{N}{r} p^r q^{N-r}$.

Abercrombie, D. (1967). *Elements of General Phonetics* (Aldine, Chicago).
 Austin, G. (1806). *Chironomia* (Cadell and Davies, London). Reprinted by Southern Illinois University Press, Carbondale, IL, 1966.
 Belkin, K., Martin, R., Kemp, S. E., and Gilbert, A. N. (1997). “Auditory pitch as a perceptual analogue to odor quality,” *Psychol. Sci.* **8**, 340–342.

Berry, D. A., Herzel, H., Titze, I. R., and Story, B. H. (1996). “Bifurcations in excised larynx experiments,” *J. Voice* **10**, 129–138.
 Childers, D. G., and Lee, C. K. (1991). “Vocal quality factors: Analysis, synthesis, and perception,” *J. Acoust. Soc. Am.* **90**, 2394–2410.
 Cicero (1948). *De Oratore*, translated by E. W. Sutton and H. Rackham (Harvard University Press, Cambridge, MA).
 Colton, R., and Estill, J. (1981). “Elements of voice quality: Perceptual, acoustic and physiologic aspects,” in *Speech and Language: Advances in Basic Research and Practice*, edited by N. J. Lass (Academic, New York), Vol. 5, pp. 311–403.
 Fairbanks, G. (1940). *Voice and Articulation Drillbook* (Harper, New York).
 Gandour, J., and Harshman, R. (1978). “Crosslanguage differences in tone perception: A multidimensional scaling investigation,” *Lang. Speech* **21**, 1–33.
 Gandour, J., Petty, S. H., and Dardarananda, R. (1988). “Perception and production of tone in aphasia,” *Brain Lang.* **35**, 201–240.
 Gerratt, B. R., and Kreiman, J. (2000). “Toward a taxonomy of nonmodal phonation,” *J. Phonetics* (in press).
 Gerratt, B. R., Kreiman, J., Antonanzas-Barroso, N., and Berke, G. S. (1993). “Comparing internal and external standards in voice quality judgments,” *J. Speech Hear. Res.* **36**, 14–20.
 Gobl, C., and Ni Chasaide, A. (1992). “Acoustic characteristics of voice quality,” *Speech Commun.* **11**, 481–490.
 Goldbury, J., and Russell, W. (1844). *The American Common-School Reader and Speaker* (Tappan and Whitmore, Boston) (cited by Gray, 1943).
 Gray, G. W. (1943). “The ‘voice qualities’ in the history of elocution,” *Q. J. Speech* **29**, 475–480.
 Hays, W. L. (1994). *Statistics*, 5th ed. (Harcourt Brace, New York).
 Herzel, H., Steinecke, I., Mende, W., and Wermke, K. (1991). “Chaos and bifurcations during voiced speech,” in *Complexity, Chaos, and Biological Evolution*, edited by I. Mosekilde and L. Mosekilde (Plenum, New York), pp. 41–50.
 Hillenbrand, J., and Houde, R. A. (1996). “Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech,” *J. Speech Hear. Res.* **39**, 311–321.
 Hollien, H., Girard, G. T., and Coleman, R. F. (1977). “Vocal fold vibratory patterns of pulse register phonation,” *Folia Phoniatr.* **29**, 200–205.
 Hollien, H., Moore, P., Wendahl, R. W., and Michel, J. (1966). “On the nature of vocal fry,” *J. Speech Hear. Res.* **9**, 245–247.
 Hollien, H., and Wendahl, R. W. (1968). “Perceptual study of vocal fry,” *J. Acoust. Soc. Am.* **43**, 506–509.
 Jones, D. (1922). *An Outline of English Phonetics*, 9th ed. (Cambridge University Press, Cambridge, 1972).
 Kasuya, H., and Ando, Y. (1991). “Acoustic analysis, synthesis, and perception of breathy voice,” in *Vocal Fold Physiology: Acoustic, Perceptual, and Physiological Aspects of Voice Mechanisms*, edited by J. Gauffin and B. Hammarberg (Singular, San Diego), pp. 251–258.
 Klatt, D. H., and Klatt, L. C. (1990). “Analysis, synthesis, and perception of voice quality variations among female and male talkers,” *J. Acoust. Soc. Am.* **87**, 820–857.
 Kreiman, J., and Gerratt, B. R. (1996). “The perceptual structure of pathologic voice quality,” *J. Acoust. Soc. Am.* **100**, 1787–1795.
 Kreiman, J., and Gerratt, B. R. (1998). “Validity of rating scale measures of voice quality,” *J. Acoust. Soc. Am.* **104**, 1598–1608.
 Kreiman, J., Gerratt, B. R., Kempster, G., Erman, A., and Berke, G. S. (1993). “Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research,” *J. Speech Hear. Res.* **36**, 21–40.
 Kreiman, J., Gerratt, B. R., and Precoda, K. (1990). “Listener experience and perception of voice quality,” *J. Speech Hear. Res.* **33**, 103–115.
 Lalwani, A. L., and Childers, D. G. (1991). “Modeling vocal disorders via formant synthesis,” *Proc. ICASSP*, pp. 505–508.
 Laver, J. (1980). “The analysis of vocal quality: From the classical period to the twentieth century,” in *Towards a History of Phonetics*, edited by R. E. Asher and E. J. A. Henderson (Edinburgh University Press, Edinburgh), pp. 79–99.
 Martin, D., Fitch, J., and Wolfe, V. (1995). “Pathologic voice type and the acoustic prediction of severity,” *J. Speech Hear. Res.* **38**, 765–771.
 Martin, D., and Wolfe, V. (1996). “Effects of perceptual training based upon synthesized voice signals,” *Percept. Mot. Skills* **83**, 1291–1298.
 Michel, J., and Hollien, H. (1968). “Perceptual differentiation of vocal fry and harshness,” *J. Speech Hear. Res.* **11**, 439–443.
 Milenkovic, P. H. (1987). “Least mean square measure of voice perturbation,” *J. Speech Hear. Res.* **30**, 529–538.

- Milenkovic, P., and Read, C. (1992). *CSpeech Version 4 User's Manual* (Paul Milenkovic, Madison, WI).
- Omori, K., Kojima, H., Kakani, R., Slavit, D. H., and Blaugrund, S. M. (1997). "Acoustic characteristics of rough voice: Subharmonics," *J. Voice* **11**, 40–47.
- Peterson, G. E., and Barney, H. L. (1952). "Control methods used in a study of the vowels," *J. Acoust. Soc. Am.* **24**, 175–184.
- Plomp, R. (1976). *Aspects of Tone Sensation* (Academic, New York).
- Plugge, D. E. (1942). "'Voice qualities' in oral interpretation," *Q. J. Speech* **28**, 442–444.
- Pollux, J. (1706). *Onomasticon*, Amsterdam edition (cited by Austin, 1806).
- Rammage, L. A., Peppard, R., and Bless, D. M. (1992). "Aerodynamic, laryngoscopic, and perceptual-acoustic characteristics in dysphonic females with posterior glottal chinks: A retrospective study," *J. Voice* **6**, 64–78.
- Rush, J. (1859). *The Philosophy of the Human Voice*, 5th ed. (J. B. Lippincott, Philadelphia).
- Wendahl, R. W. (1966). "Some parameters of auditory roughness," *Folia Phoniatr.* **18**, 26–32.