The perceptual structure of pathologic voice quality

Jody Kreiman and Bruce R. Gerratt
Division of Head and Neck Surgery, UCLA School of Medicine, 31-24 Rehab Center, Los Angeles, California 90095-1794

(Received 5 July 1995; revised 2 February 1996; accepted 3 April 1996)

Although perceptual assessment is included in most protocols for evaluating pathologic voices, a standard set of valid scales for measuring voice quality has never been established. Standardization is important for theory and for clinical acceptance, and also because validation of objective measures of voice depends on valid perceptual measures. The present study used large sets ($n=80$) of male and female voices, representing a broad range of diagnoses and vocal severities. Eight experts judged the dissimilarity of each pair of voices, and responses were analyzed using nonmetric individual differences multidimensional scaling. Results indicate that differences between listeners in perceptual strategy are so great that the fundamental assumption of a common perceptual space must be questioned. Because standardization depends on the assumption that listeners are similar, it is concluded that efforts to standardize perceptual labels for voice quality are unlikely to succeed. However, analysis by synthesis may provide an alternate means of modeling quality as a function of both voices and listeners, thus avoiding this problem. © *1996 Acoustical Society of America.*

PACS numbers: 43.71.Bp, 43.71.Gv, 43.70.Dn [RAF]

INTRODUCTION

Most protocols for evaluating pathologic voices include perceptual assessment of quality (e.g., Greene and Mathieson, 1989; Hirano, 1989; Gerratt *et al.*, 1991; American Speech-Language-Hearing Association, 1993). Various instruments have been proposed for these perceptual evaluations, including the GRBAS scales (Hirano, 1981), the Wilson voice profile (Wilson, 1977), Laver's voice profile analysis (Wirz and Beck, 1995), and the scales described by Hammarberg and Gauffin (1995). However, none of these protocols has been widely accepted, either clinically or for research purposes, because of ongoing concerns about the validity of the scales used to assess quality, and about the reliability with which listeners can rate or label voices.

Concerns about scale validity derive from the fact that researchers have never established what features of pathologic voices are perceptually important. Without this knowledge, studies examining listener reliability may be difficult to interpret. That is, when listeners disagree, or when training protocols do not improve listener reliability, results may reflect listener vagaries, or the fact that the scales or training stimuli do not adequately represent voice quality, and thus do not measure what listeners hear.

Although validity is of primary importance in any measurement system, voice researchers have traditionally neglected this issue, and have concentrated more on rater reliability (see Kreiman *et al.*, 1993, for review). Only a few studies have formally investigated the perceptually important attributes of pathologic voices, using factor analysis (FA) (Isshiki and Takeuchi, 1970; Takahashi and Koike, 1975; Hammarberg *et al.*, 1980, 1986; Nieboer *et al.*, 1988) or multidimensional scaling (MDS) (Murry *et al.*, 1977; Kreiman *et al.*, 1990, 1992, 1994; Kempster *et al.*, 1991). Unfortunately, methodological difficulties limit the conclusions that can be drawn from these studies. Most studies have used rather small samples of voices ($n=15$ to 31). This is a significant limitation, because results of both FA and MDS depend on the sample of voices: Qualities can only emerge as perceptually important if they are represented in the set of voices being evaluated. Results of different studies have reflected differences in the samples used, and only a single dimension ($F0$) is common to all MDS studies (Kreiman *et al.*, 1990). Further, the small number of voices studied has limited the number of dimensions that may be extracted from a set of data, so that half or more of the variance in the underlying ratings may remain unexplained (Murry *et al.*, 1977; Nieboer *et al.*, 1988).

Other sampling restrictions further limit the generalizability of previous studies. Many published reports have used normal voices (e.g., Matsumoto *et al.*, 1973; Walden *et al.*, 1978; Murry and Singh, 1980; Gelfer, 1993), but pathologic voices have been studied very little. Many studies that did examine pathologic voices included only a single disordered population [e.g., tracheoesophageal speakers (Nieboer *et al.*, 1988), hoarse voices (Isshiki and Takeuchi, 1970; an exception is Hammarberg *et al.*, 1980] or focused on specific aspects of voice quality (Kreiman *et al.*, 1994). Others included significant numbers of normal voices (e.g., Takahashi and Koike, 1975), reducing their sensitivity to differences among pathologic qualities. Factor analysis has the further limitation that results depend on the set of scales studied, as well as on the set of voices. If scales related to some important quality are omitted, that quality will not be reflected in the results.

Finally, relatively little has been done to examine differences among listeners in the perceptual strategies they use when evaluating pathologic voices. Such differences are of considerable theoretical importance, because attempts to standardize terminology or develop protocols for perceptual assessment depend on the assumption that listeners share common features for pathologic vocal quality. Specifically, we assume that although listeners may differ substantially in

the importance they assign to different perceptual dimensions, the perceptually real dimensions themselves do not vary widely from listener to listener. This assumption is critical: If listeners differ substantially in the perceptual strategies they apply when evaluating voices, then a single valid, nonarbitrary set of scales for quality cannot be defined.

Given the assumption that listeners are essentially interchangeable, quality can be modeled as an attribute of voices, rather than as a unique interaction between a given listener and a particular voice. Of course, like pitch or loudness, voice quality actually is an interaction between voices and listeners: Quality is not an attribute of voices apart from the listener's act of hearing, any more than pitch is an attribute of signals. However, researchers traditionally have made the simplifying assumption that a consistent mapping exists between acoustic signals and perceived quality, analogous to a psychophysical function relating pitch to frequency. Given this assumption, it is further possible to assume that a ''right answer'' does exist for quality judgments. That is, if (and only if) listeners are essentially the same in the way they process the relevant acoustic information, it follows that quality, however complex, *can* be assessed directly as a function of the voice signal, and that it is reasonable to expect listeners to agree in their assessments of quality, as they do in their assessments of pitch or loudness. From this perspective, failures of agreement in listener ratings of voice are noise, which must be controlled to recover this ''right answer.'' Alternatively, failure to agree in ratings may be interpreted as evidence of fundamental differences in the way listeners process the acoustic signal.

The present study used MDS to examine the perceptual structure of large samples of pathologic male and female voices, in an attempt to establish a valid set of perceptual features for pathologic vocal quality that will generalize to any clinical setting or population. MDS has provided insight into the perceptual structure of complex stimuli in several sensory domains (see Schiffman *et al.*, 1981; Jones and Koehly, 1993, for review), and has been useful in accounting for listeners' perceptions of pathological voice (Murry *et al.*, 1977; Kreiman *et al.*, 1990, 1992, 1994; Kempster *et al.*, 1991), with the limitations noted above. Because listeners judge the *dissimilarity* of pairs of voices, no assumptions are required about what scales or features might be needed to describe quality. Results are generally considered to be ''psychologically real,'' in that they reflect the perceptual structure of the stimulus set. Results may reflect uninteresting aspects of the stimulus set; for example, if the voices studied differ widely in $F0$, these differences may dominate listeners' similarity ratings, at the expense of more subtle qualities. However, this limitation can be circumvented by using very large sets of stimuli so that features such as pitch and loudness—or any others—may emerge, along with more traditional ''qualities,'' if listeners consistently attend to this information. Finally, MDS permits investigation of differences among listeners in perceptual strategy.

I. METHODS

A. Speakers and voice samples

Two sets of voices were selected from a large library of recordings made under identical conditions as part of a phonatory function analysis. One set included 80 female voices, the other 80 male voices. Stable MDS solutions may be achieved with as few as six stimuli for each hypothetical dimension (Schiffman *et al.*, 1981). Thus these samples are large enough to support solutions with as many as 13 dimensions.

Voices were recorded using a microphone placed off-axis 5 cm away from the speaker's lips. Speakers were asked to sustain /ɑ/ for as long as possible. Utterances were low-pass filtered at 8 kHz and sampled at 20 kHz with 12-bit resolution. A 2-s sample was excerpted from the middle of each utterance and stored for later presentation. Stimuli were equalized for peak intensity, and onsets and offsets were multiplied by 40-ms ramps to eliminate click artifacts.

Steady-state vowels, rather than continuous speech, were studied for several reasons. Listeners' ratings of quality from vowels and connected speech are similar (e.g., de Krom, 1995); relatively short stimuli enabled us to gather ratings for larger sets of voices; and we hoped that the vowels' relatively simple acoustic structure would yield more consistent, interpretable perceptual strategies than would more complex continuous speech. A brief vowel may not represent the full range of qualities produced by a particular speaker. However, across speakers the spectrum of possible vocal qualities was well represented, as argued below. Study of continuous speech is an obvious next step once valid results based on less complex stimuli are obtained.

To ensure that a broad range of vocal qualities, underlying pathologies, and severities were represented, voice selection took place in several steps. First, the entire library of over 1000 recordings was reviewed. Each voice was given a ''severity'' rating by consensus vote of the authors and an experienced speech-language pathologist, with disagreements resolved by discussion. Severity ratings were made on a six-point equal-appearing interval (EAI) scale, where 1 represented near-normal voice quality and 6 extremely severe pathology. To avoid biasing results, voices were unselected with respect to any particular qualities. Each speaker's age and diagnosis were obtained from medical records. Female speakers ranged in age from 22 to 89 years (mean $=60.0$; s.d.$=17.9$). Male speakers ranged in age from 18 to 96 years (mean$=52.8$; s.d.$=16.5$). Diagnoses were divided into six categories: mass lesion; paralysis, paresis or glottal gap; adductory spasmodic dysphonia; functional disorder or vocal abuse; neuromuscular disorder; and ''other'' (including ideopathic disorders, trauma, hemilaryngectomy, and atrophy). Voices were deleted from each set in such a fashion that a range of ages and severities was present for each diagnosis, and such that each diagnostic category was about equally represented. The final set of voices included roughly equal numbers of mildly, moderately, and severely pathologic voices in each diagnostic category. Chi-square analysis showed no asymmetries in the distribution of samples by diagnosis and severity of pathology (females: $\chi^2=21.11$; df

$=25$; $p>0.05$; males: $\chi^2=16.87$; $df=25$; $p>0.05$). A two-way ANOVA confirmed that the speakers' ages did not vary significantly with gender [$F(1,140)=0.01$, $p>0.05$] or diagnosis [$F(5,140)=1.98$, $p>0.05$]; and no gender by diagnosis interaction occurred [$F(5,140)=0.48$, $p>0.05$]. Because severity of pathology, diagnostic category, and the speakers' ages and gender are unconfounded in these voice sets, we assume that the populations of interest are adequately sampled.

B. Listeners

Twelve expert listeners (six speech-language pathologists, five otolaryngologists, and one phonetician, including both authors) participated in these experiments. Four listeners (two speech-language pathologists, one otolaryngologist, and one phonetician) heard both voice sets; four (two speech-language pathologists and two otolaryngologists) heard only the female voices, and four (two speech-language pathologists and two otolaryngologists) heard only the male voices, for a total of eight listeners/voice set. Each listener had a minimum of three years' experience evaluating and/or treating voice disorders. Listeners reported no history of any hearing, speech, voice, or language difficulties. They were screened for ability to detect pure tones at 25 dB HL at octave frequencies from 500 Hz to 4 kHz.

C. Procedure

For each voice set, listeners heard one order (AB or BA) of each pair of voices (3160 comparisons/listener/voice set). Voices within a pair were separated by 0.5 s. Which voice occurred first within a pair varied at random, with the constraint that each voice occurred first an equal number of times. An additional 632 pairs (20%), selected at random, were repeated for each voice set so that test–retest reliability could be assessed. Order of stimulus presentation was randomized across listeners and test sessions; repeated pairs were inserted at random into the total voice set, with the constraint that identical pairs of voices were separated by at least ten pairs. Test time for a single listener totalled about 20 h. To minimize listener fatigue, testing took place in 1-h sessions, one per week for 20 weeks.

Listeners judged the dissimilarity of each pair of voices on a seven-point EAI scale, where ''1'' meant the two voices were extremely similar in quality, and ''7'' meant they were extremely different. The rate of presentation was controlled by the listeners, who were able to replay the pairs as necessary before responding. They were asked to concentrate and to avoid labeling the qualities a voice might have or guessing the diagnoses.

TABLE I. Intrarater reliability and agreement.

	% ratings +/−1 scale value		r for 1st vs 2nd ratings	
Voice set	mean	range	mean	range
Females	77.3%	68.1%–83.9%	0.65	0.50–0.76
Males	74.4%	60.6%–82.1%	0.63	0.40–0.78

Listeners were tested individually. To mimic clinical listening conditions, all testing took place in free field. Listeners were seated 3 ft from a speaker in a sound-treated room. Stimuli were played through a 16-bit D/A converter, reconstruction filtered at 8 kHz, and presented at a constant listening level (approx. 80 dB SPL). Responses were recorded and stored by the computer.

II. RESULTS

A. Reliability and agreement

Intrarater reliability was impressive, particularly given the long duration of testing (Table I). On the average, 75% of repeated ratings were within ± 1 scale value (chance $=38.8\%$); more than 60% of ratings met this criterion for even the most unreliable listeners. Values of Pearson's r are somewhat lower, probably due to non-normal rating distributions and/or limited use of one or both scale extremes by some listeners.

Analyses of interrater reliability showed the typical pattern (e.g., Gerratt *et al.*, 1993; Kreiman *et al.*, 1993) of moderate average agreement among listeners with large differences in the agreement levels observed for particular pairs of listeners (Table II). Again consistent with previous studies, the ratings received by individual voice pairs were highly variable. Across voice sets, only 23 pairs (0.4%) received the same rating from all eight raters. One hundred thirty-one pairs (2.1%) received ratings spanning the entire seven-point scale, and an additional 765 pairs (12.1%) received ratings spanning six of the seven points (one–six or two–seven). Values of the intraclass correlation reflect this variability, with an average of only 24% of variance common to the eight listeners in each group.

B. MDS analyses for the group data

Listeners' judgments were analyzed using a nonmetric individual differences multidimensional scaling model (SAS Institute, Inc., 1992). Separate solutions in 1 to 6 dimensions were found for the female and male voice sets. Based on plots of stress and variance accounted for versus the number of dimensions extracted (e.g., Schiffman *et al.*, 1981; Fig. 1), two-dimensional solutions were selected for both voice sets.

TABLE II. Interrater reliability.

		% ratings +/−1 scale value		r for pairs of raters	
Voice set	ICC	mean	range	mean	range
Females	0.50	65.2%	47.1%–77.0%	0.54	0.39–0.65
Males	0.47	59.5%	47.6%–68.9%	0.50	0.36–0.61

Note: ICC=Intraclass correlation.

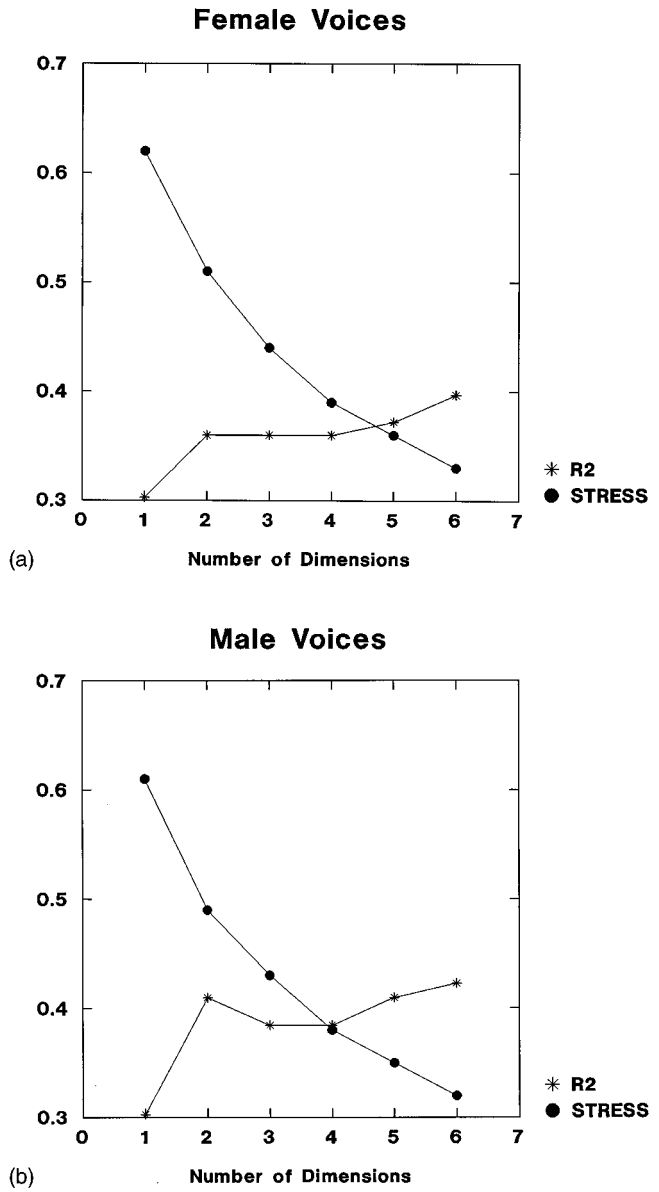

FIG. 1. Values of stress (filled circles) and variance accounted for (r^2; stars) for multidimensional scaling solutions of group data. (a) Female voices. (b) Male voices.

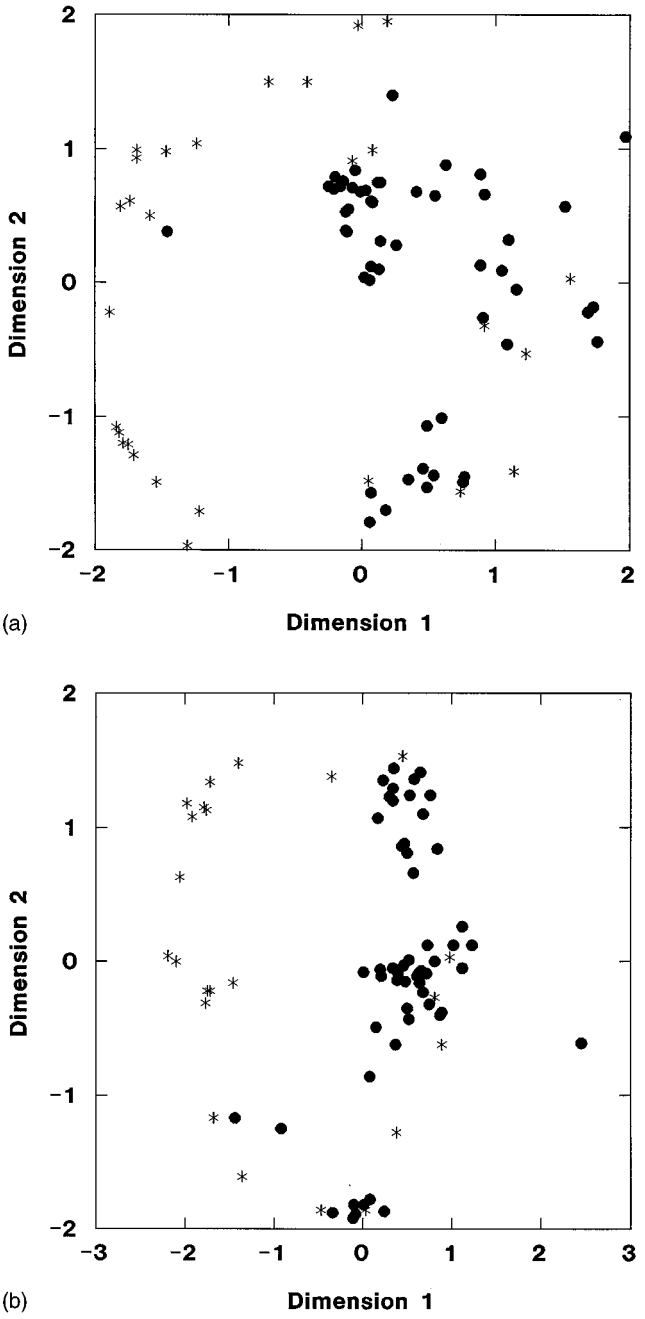

FIG. 2. Perceptual spaces for group data. Stars represent severely pathologic voices; filled circles represent voices with mild to moderate pathology. (a) Female voices. (b) Male voices.

For both solutions, r^2 values were quite low (female voices: $r^2=0.36$; male voices: $r^2=0.41$), and stress values were quite high (female voices; stress=0.51; male voices: stress=0.49).

The group perceptual spaces are shown in Fig. 2. The first dimension in the spaces for both female and male voices separated stimuli into two groups: those with severity ratings of five or six (shown as stars in the figure), versus those with severity ratings of four or less [filled circles; female voices: χ^2 (1)=43.63; male voices: χ^2 (1)=38.36, $p<0.05$]. The second dimension in both spaces appears to divide the voices into clusters, but subjectively these lacked unifying percepts. Attempts at using regression, correlation, or discriminant analysis to interpret the spaces in terms of acoustic characteristics of the stimuli were unsuccessful, due both to the clustered nature of the spaces and to difficulties in making acoustic measurements for many aperiodic samples.

C. MDS analyses for individual listeners

Group MDS analyses assume a common perceptual space, from which listeners may deviate in the weight they give to particular dimensions, but not in the dimensions they perceive. The high values of stress, low amounts of variance accounted for, and the low dimensionality of the group perceptual spaces probably reflect violations of this assumption. Further, across voice sets, only 134/6320 voice pairs (2.1%) had mean ratings of two or less (quality highly similar); 1095/6320 voice pairs (17.3%) had mean ratings of six or more, and nearly half the voice pairs (3140/6320;49.7%) re-

ceived a rating of seven (quality very different) from at least one rater. This also suggests that listeners lacked a common sense of ''similarity.''

To examine this hypothesis, separate MDS analyses were undertaken for each listener for each voice set. For the female voices, three dimensional solutions were selected for four listeners and two-dimensional solutions for the other four; stress ranged from 0.27–0.36, and r^2 ranged from 0.56–0.83. For the male voices, two three-dimensional and six two-dimensional solutions were selected; stress ranged from 0.27–0.33, and r^2 ranged from 0.58–0.77.

Representative individual scaling solutions for male voices are shown in Fig. 3. Three solutions resembled the space in Fig. 3(a), consisting of a central cluster or clusters of mildly to moderately pathologic voices, surrounded by an arc of severely pathologic stimuli. Listeners differed in the number of voices in the arc ($n=26$–43), but ordering of common voices ($n=16$) was consistent across listeners (Spearman's rho=0.92–0.99). The perceptual space for a fourth listener divided voices into the same mild and severe groups, with the severely pathologic voices arranged around the central cluster of mild-to-moderately pathologic stimuli. However, the peripheral voices in this solution were arranged in clusters, rather than continuously, and their ordering did not correspond to that of the other three listeners (Spearman's rho=0.30–0.32).

Perceptual spaces for the remaining four listeners consisted of clusters of voices [Fig. 3(b)]. Listeners differed considerably in the nature of those groupings, although spaces for three of the four listeners included a primary distinction between extreme pathology and less severe dysphonias, with either the more severe (two listeners) or the milder samples (one listener) clustering together. In the fourth space, no relationship between severity and clustering was detected.

Typical spaces for female voices are shown in Fig. 4. Spaces for four listeners included combinations of continua and clusters of voices [Fig. 4(a)]. The four other spaces consisted entirely of clusters [Fig. 4(b)]. Samples with milder pathology grouped together for five listeners, while severely pathologic samples formed a cluster for one other. No relationship between severity of pathology and clustering was observed for the remaining two listeners.

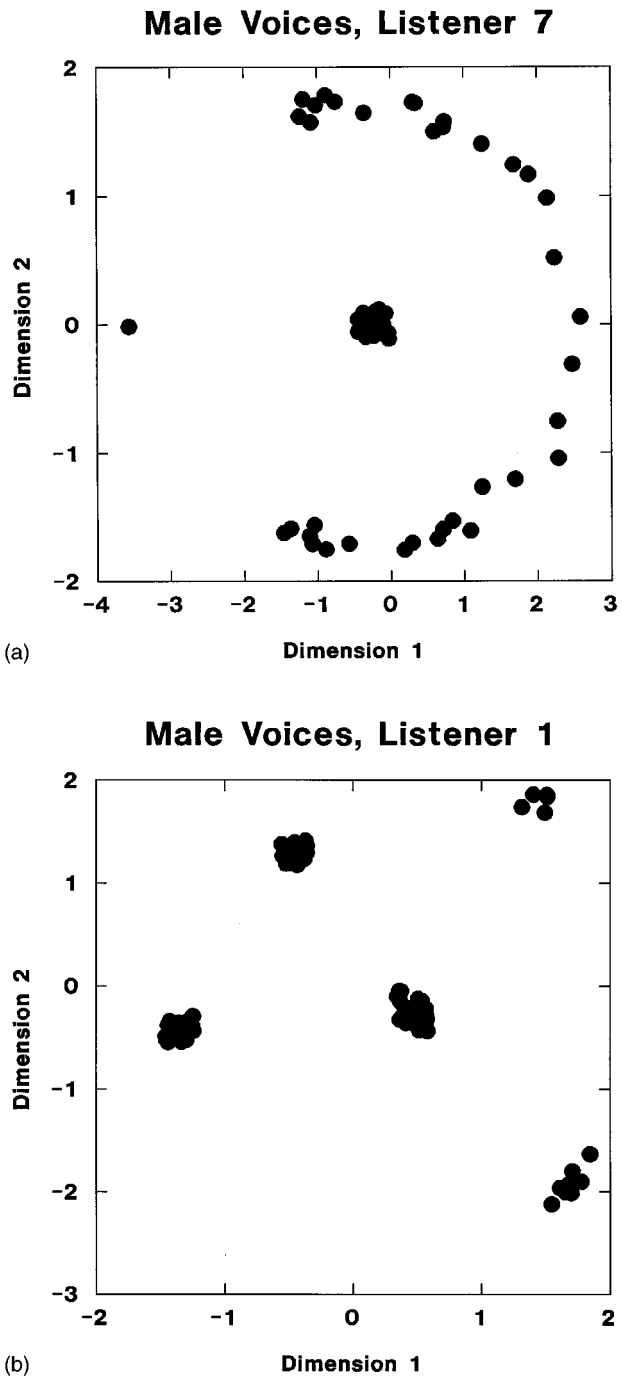

FIG. 3. Representative perceptual spaces for individual listeners: Male voice set.

D. Cluster analyses

As for the group perceptual spaces, efforts to use regression, correlation, or discriminant analysis to interpret dimensions in these individual perceptual spaces in terms of measured characteristics of the signals were unsuccessful. Interpretation of the dimensions in terms of rated characteristics of the voices was attempted informally but not pursued, because of the large differences observed between listeners in voice rating (cf. Kreiman *et al.*, 1993), and because of the concerns regarding scale validity that originally motivated this study. We attempted to give the clusters impressionistic labels, but found that the qualities represented often overlapped. All the voices in a cluster sounded ''similar'' to the others in one way or another, but more than one ''quality'' was generally represented in a cluster and other clusters also included voices with similar qualities.

To quantify the extent to which clusters represented stable categories of voices across listeners, without resorting to impressionistic ratings or labels, we determined how often any two voices were placed together in a cluster. If voices are consistently placed together, then we can conclude that they share some salient ''feature'' or that they represent a relatively stable perceptual category.

Clusters in each perceptual space were defined objec-

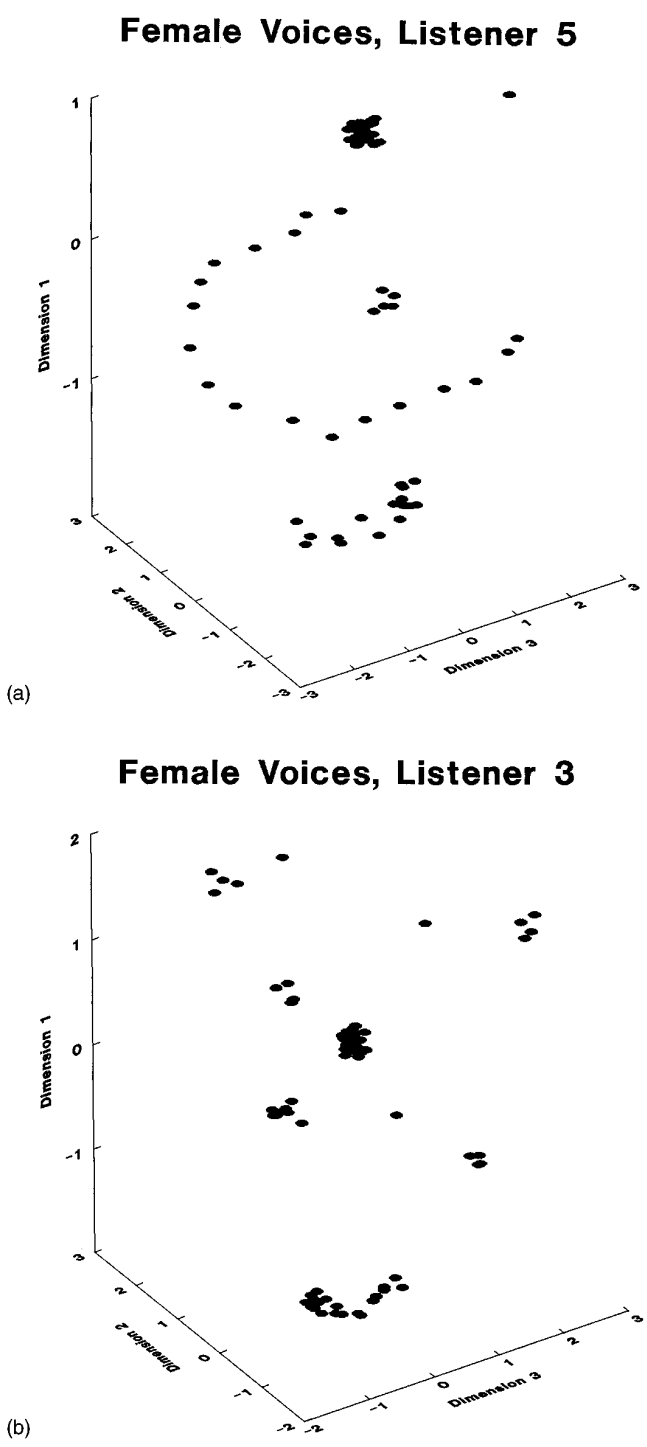

FIG. 4. Representative perceptual spaces for individual listeners: Female voice set.

tively with *k*-means cluster analysis. Pairs of listeners placed the voices in an average of 10.6% of pairs in the same cluster (range=5.4% to 27.4%). However, no two voices were placed in the same cluster by all eight listeners, for either voice set; for 0.7% of voice pairs, seven listeners placed the voices in the same cluster, and for 2.4% of pairs, the voices were placed in the same cluster by six listeners. For 14.1% of pairs, those two voices were placed in different clusters by all eight listeners; for 20.4% of pairs, the voices were placed in different clusters by seven listeners, and for 21.7% of pairs, the voices were placed in different clusters by six listeners.

E. Effects of sample size on MDS solutions

Previous MDS studies of pathologic voice quality (e.g., Murry *et al.*, 1977; Kreiman *et al.*, 1990, 1992; Kempster *et al.*, 1991) have revealed meaningful dimensions, albeit different dimensions from each study (Kreiman *et al.*, 1990). However, these previous studies used rather small sets of voices ($n=15$ to 31). Given the perceptual complexity of pathologic voices, these positive but variable findings may reflect sampling error. We speculated that meaningful dimensions would emerge from virtually any small set of voices, because there are a limited number of judgments involved in such analyses and thus a limit to the complexity of the perceptual relations that are possible.

To test this sampling error hypothesis, four 15-voice random samples were drawn from the 80 voice sets (two from the male set and two from the female set). The group MDS analysis (nonmetric individual differences model) was repeated for each random sample, using only the dissimilarity ratings for pairs of the 15 voices in that sample. Each analysis included eight 15×15 lower half matrices, one for each listener.

Results are summarized in Table III. Two-dimensional solutions were selected for both samples of female voices; for the male voices, one two-dimensional solution and one one-dimensional solution were selected. These solutions accounted for virtually all the variance in the underlying data. Each solution was interpretable in terms of the acoustic characteristics of the voices (see Kreiman *et al.*, 1994, for details of the acoustic analyses). However, different interpretations were obtained for each solution, and correlations between stimulus coordinates in the new and old solution spaces were low.

These findings are consistent with the view that previous reports of ''dimensions'' for pathologic voice quality reflect sampling error. Given a small set of voices, ''features'' may be derived that will account for the patterns of similarity among those voices, and interpretable perceptual structure will emerge. However, as the sample size increases, perceptual relations become too complicated to be summarized with dimensions or features. Thus findings from these small samples do not generalize well to the underlying population of voices, or to other small samples.

V. DISCUSSION

Certain limitations of the present study must be noted. Relatively short, simple stimuli were used, which presumably limited the complexity of the perceptual strategies listeners employed; and information about nuances in voice quality may have been lost through use of a dissimilarity judgment paradigm, where a single number summarizes the overall relationship between two stimuli. Nevertheless, even given the possible reduction in available information, observed differences among listeners in perceptual strategy were so great that the fundamental assumption of a common perceptual space for pathologic voice quality must be ques-

TABLE III. Scaling solutions for 15 voice random samples drawn from the original 80 voice sets.

Voice set	Random sample #	r^2 for solution	Interpretation	Correlation with the original space
Females	1	0.86	$D1$: RPK (r^2=0.44)	$D1$: R=0.14 (ns)
			$D2$: tremor+HNR s.d. (R^2=0.41)	$D2$: R=0.44 (ns)
	2	0.90	$D1$: age+severity (R^2=0.70)	$D1$: R=0.39 (ns)
			$D2$: $F0$ (r^2=0.52)	$D2$: R=0.65 ($p<0.05$)
Males	1	0.90	$D1$: mean jitter (r^2=0.67)	$D1$: R=0.44 (ns)
			$D2$: HNR s.d. (r^2=0.66)	$D2$: R=0.20 (ns)
	2	0.86	$D1$: HNR (r^2=0.37)	$D1$: R=0.65 ($p<0.05$)

Note: RPK=Pearson's correlation between a signal and a delayed copy at a delay corresponding to the peak in the autocorrelation function (Hillenbrand *et al.*, 1994); HNR=harmonics to noise ratio (Yumoto *et al.*, 1982). See Kreiman *et al.*, 1994, for details of these analyses.

tioned. Although individual listeners were quite reliable in judging the similarity of the stimuli, perceived similarity was not constant across listeners, and features did not emerge that predicted which voices sounded similar, either across the group or within single listeners. Only a single dimension (severe versus mild/moderate pathology, possibly corresponding to the ''Grade'' scale) consistently emerged from group and individual scaling solutions. Perceptual spaces for individual listeners were characterized not by continuous dimensions, but by clusters of voices that lacked significant acoustic correlates or subjective unifying characteristics. No consistent clustering of voices was observed, so no ''features'' or ''categories'' for voices could be derived (other than the distinction between severe and mild/moderate pathology). Instead, as a group the spaces seemed structured by ''family resemblances'' (Wittgenstein, 1958; Lindau, 1985): Each voice in a cluster resembled others with respect to some property, but no one property linked all the voices within a cluster. Further, voices that clustered together for one listener appeared in different clusters for others. Listeners sometimes agreed about which voices belonged in separate clusters, but never about which voices grouped together. It appears that listener dependencies are not a negligible part of quality, and thus that quality cannot be treated solely as an attribute of voices.

Variation among listeners is not a new finding (cf., Kubawara and Ohgushi, 1984; Kreiman *et al.*, 1992), or an unexpected one. Given the complexity of voice signals, it is not surprising that listeners should differ in how they focus their attention or in the salience they accord to different voice features. Traditionally these differences have been treated as noise by investigators, and have been ''controlled'' or ignored. Listener training (Bassich and Ludlow, 1986; Hammarberg and Gauffin, 1995) and rating protocols using anchor stimuli (Gerratt *et al.*, 1993) have been investigated as ways to increase agreement among listeners in their judgments of different qualities. Training protocols attempt to ''standardize'' different listeners' varying internal definitions for a given quality, while anchored paradigms are designed to replace unstable and idiosyncratic internal standards with constant external reference stimuli. However, both approaches assume that it is reasonable to expect listeners to agree with each other. Thus the present results suggest that such techniques will not be fruitful, because valid, nonarbitrary scales, anchors, or training procedures cannot be defined or implemented when listeners do not agree about perceptual features for voices. Our results suggest that the issues surrounding rating unreliability may be theoretically unresolvable, as long as quality is treated solely as an attribute of voices rather than as an interaction between listeners and voices.

If perceptual protocols cannot be standardized, more is compromised than the perceptual assessment of voice. Efforts to develop objective measures of voice also depend critically on valid perceptual measures, because perceptual measures validate the objective measures of pathologic voice quality. Objective measures have long been popular in research applications, and are increasingly accepted as an important tool for documenting vocal quality (e.g., Hirano, 1989; Bless, 1991; Titze, 1992; Laver *et al.*, 1992; American Speech-Language-Hearing Association, 1993). However, the *utility* of such measures depends on their consistent correspondence to physiological states or to what listeners hear. In other words, a list of acoustic (or other) voice measures is not particularly useful clinically or theoretically, apart from the information such a list provides about the laryngeal condition underlying the voice problem or a listener's perception of the vocal deviation.

If one voice signal generates more than one perceptual response (as occurred here), studies seeking reliable correlations between objective and perceptual measures will never produce consistent, replicable associations, because unique associations do not exist. Much of the voice literature represents repeated failed efforts to identify such associations. For example, many believe that jitter has consistent perceptual correlates (roughness is the most frequently mentioned; Gerratt and Kreiman, 1995), and thus that it is a good measure of vocal quality. However, correlations between measures of jitter and perceived roughness have ranged from −0.01 (Eskenazi *et al.*, 1990) to 0.98 (Heiberger and Horii, 1982); and jitter has been significantly (if not always highly) associated with many qualities, including grade/severity of pathology (Askenfelt and Hammarberg, 1986; Wolfe and Ratusnik, 1988), breathiness (Hirano *et al.*, 1988; Eskenazi *et al.*, 1990; Feijoo and Hernandez, 1990), hoarseness (Haji *et al.*, 1986; Horiguchi *et al.*, 1987), and strain (Arends *et al.*, 1990).

The present results suggest that relations between perceptual and acoustic measures of voice can be identified only by integrating acoustic and perceptual approaches to voice

quality—that is, by modeling quality as a function of both a particular signal and a specific listener. Analysis by synthesis techniques may make such modeling possible, by providing listeners with the opportunity to construct a synthetic signal that perceptually matches the natural pathologic voice under observation. Such an approach does not control for listener differences. Instead, it treats the relationship between a particular voice and listener on a particular occasion as a unique relation, and then models that relation. Given a suitable system for registering what they hear, listeners should converge on an acoustic representation of the signal that generated their particular perception, because a single acoustic signal gave rise to that perception, whatever it may be. Listeners' perceptions would thus be defined objectively by the values of the synthesizer parameters they selected to match a given natural disordered voice.

In addition to potentially providing valid parametric representations of voice quality, an analysis by synthesis approach to measuring quality may resolve two other longstanding problems in voice research. First, this approach may alleviate concerns about the validity and utility of acoustic measurements, because the analysis by synthesis process directly links acoustic signals to perceived qualities, without the addition of an intervening rating process. This confirmatory approach to validating acoustic measures with voice quality perception contrasts sharply with the purely exploratory, correlative approaches typically found in the literature. Second, analysis by synthesis techniques should increase measurement reliability, because they control several major sources of listener variability in voice quality ratings. We have argued (Gerratt *et al.*, 1993; Kreiman *et al.*, 1993) that traditional quality rating methods involve comparing a voice to an unstable, unobservable internal standard. The sample of voices with which a voice is judged, experience listening to pathological voice quality, and severity of vocal deviation can all influence this internal representation, thereby increasing rating variability. In contrast, analysis by synthesis approaches require matching two physical signals, without reference to unobservable internal standards for a particular quality or qualities. Because analysis by synthesis does not involve reference to mutable internal standards, we hypothesize that it will prove more reliable than traditional rating paradigms for vocal quality.

In conclusion, we believe that evaluation of pathologic voices with rating scales will never be adequate for clinical or experimental purposes, because valid rating scales do not appear to be definable so that they specify vocal quality across listeners. However, logical and theoretical considerations suggest that analysis by synthesis techniques may provide a tool for modeling quality in a way that is theoretically attractive. Modeling pathologic vocal quality as a function of both listeners and voices greatly complicates the measurement of voice quality, and many improvements in synthesis of pathologic voices will be necessary before such a protocol can be implemented (Alwan *et al.*, 1995). Converging evidence from other paradigms is also required before traditional voice rating paradigms can be abandoned. Nevertheless, the present results cannot be accommodated by traditional views of voice quality, and suggest that new approaches are necessary if quality is ever to be validly and reliably measured.

ACKNOWLEDGMENTS

We thank our expert listeners Steven Bielamowicz, Ken Briskin, Andrew Erman, Jeanne Katzman, Sina Nasri, Julie Pardo, C. Rose Rabinov, Ian Storper, Julie Trautman, and Mary Walsh for their patience and dedication. Norma Antonanzas-Barroso provided stimulus presentation and analysis software, and Christina Foreman performed many analyses of the data. Britta Hammarberg, Minoru Hirano, and one anonymous reviewer provided many helpful comments on an earlier version of this paper. This research was supported by grant DC 01797 from the National Institute on Deafness and Other Communication Disorders and by Veterans Administration Merit Review funds, and was carried out at the VA Medical Center, West Los Angeles.

Alwan, A., Bangayan, P., Kreiman, J., and Long, C. (**1995**). ''Time and frequency synthesis parameters for severe pathological voice qualities,'' in *Proceedings of the International Congress of Phonetic Sciences, Stockholm.*

American Speech-Language-Hearing Association (**1993**). ''Preferred practice patterns for the professions of speech-language pathology and audiology: Voice assessment,'' ASHA **35**, Suppl. 1, 69–70.

Arends, N., Povel, D., Os, E. von, and Speth, L. (**1990**). ''Predicting voice quality of deaf speakers on the basis of glottal characteristics,'' J. Speech Hear. Res. **33**, 116–122.

Askenfelt, A. G., and Hammarberg, B. (**1986**). ''Speech waveform perturbation analysis: A perceptual-acoustical comparison of seven measures,'' J. Speech Hear. Res. **29**, 50–64.

Bassich, C., and Ludlow, C. L. (**1986**). ''The use of perceptual methods by new clinicians for assessing voice quality,'' J. Speech Hear. Disord. **51**, 125–133.

Bless, D. M. (**1991**). ''Measurement of vocal function,'' Otolaryngologic Clin. North Am. **24**, 1023–1033.

de Krom, G. (**1995**). ''Some spectral correlates of pathological breathy and rough voice quality for different types of vowel fragments,'' J. Speech Hear. Res. **38**, 794–811.

Eskenazi, L., Childers, D. G., and Hicks, D. M. (**1990**). ''Acoustic correlates of vocal quality,'' J. Speech Hear. Res. **33**, 298–306.

Feijoo, S., and Hernandez, C. (**1990**). ''Short-term stability measures for the evaluation of vocal quality,'' J. Speech Hear. Res. **33**, 324–334.

Gelfer, M. P. (**1993**). ''A multidimensional scaling study of voice quality in females,'' Phonetica **50**, 15–27.

Gerratt, B. R., and Kreiman, J. (**1995**). ''Utility of acoustic measures of voice,'' in *Proceedings of the Workshop on Standardization in Acoustic Voice Analysis*, edited by D. Wong (Denver Center for the Performing Arts, Denver), pp. GER-1–GER-7.

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.

Gerratt, B. R., Till, J., Rosenbek, J. C., Wertz, R. T., and Boysen, A. E. (**1991**). ''Use and perceived value of perceptual and instrumental measures in dysarthria management,'' in *Dysarthria and Apraxia of Speech*, edited by C. A. Moore, K. M. Yorkston, and D. R. Beukelman (Brookes, Baltimore), pp. 77–93.

Greene, M., and Mathieson, L. (**1989**). *The Voice and Its Disorders* (Singular, San Diego).

Haji, T., Horiguchi, S., Baer, T., and Gould, W. J. (**1986**). ''Frequency and amplitude perturbation analysis of electroglottograph during sustained phonation,'' J. Acoust. Soc. Am. **80**, 58–62.

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

Hammarberg, B., Fritzell, B., Gauffin, J., and Sundberg, J. (**1986**). ''Acoustic and perceptual analysis of vocal dysfunction,'' J. Phon. **14**, 533–547.

Hammarberg, B., and Gauffin, J. (**1995**). ''Perceptual and acoustic characteristics of quality differences in pathological voices as related to physi-

ological aspects,'' in *Vocal Fold Physiology: Voice Quality Control*, edited by O. Fujimura and M. Hirano (Singular, San Diego), pp. 283–303.
Heiberger, V. L., and Horii, Y. (**1982**). ''Jitter and shimmer in sustained phonation,'' in *Speech and Language: Advances in Basic Research and Practice*, edited by N. J. Lass (Academic, New York), Vol. 7, pp. 299–332.
Hillenbrand, J., Cleveland, R. A., and Erickson, R. L. (**1994**). ''Acoustic correlates of breathy vocal quality,'' J. Speech Hear. Res. **37**, 769–778.
Hirano, M. (**1981**). *Clinical Examination of Voice* (Springer-Verlag, New York).
Hirano, M. (**1989**). ''Objective evaluation of the human voice: Clinical aspects,'' Folia Phoniatr. **41**, 89–144.
Hirano, M., Hibi, S., Yoshida, T., Hirade, Y., Kasuya, H., and Kikuchi, Y. (**1988**). ''Acoustic analysis of pathological voice,'' Acta Otolaryngolog. (Stockholm) **105**, 432–438.
Horiguchi, S., Haji, T., Baer, T., and Gould, W. J. (**1987**). ''Comparison of electroglottographic and acoustic waveform perturbation measures,'' in *Vocal Fold Physiology: Laryngeal Function in Phonation and Respiration*, edited by T. Baer, C. Sasaki, and K. Harris (College Hill, San Diego), pp. 509–518.
Isshiki, N., and Takeuchi, Y. (**1970**). ''Factor analysis of hoarseness,'' Stud. Phonolog. **5**, 37–44.
Jones, L. E., and Koehly, L. M. (**1993**). ''Multidimensional scaling,'' in *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues*, edited by G. Keren and C. Lewis (Erlbaum, Hillsdale, NJ), pp. 95–164.
Kempster, G. B., Kistler, D. J., and Hillenbrand, J. (**1991**). ''Multidimensional scaling analysis of dysphonia in two speaker groups,'' J. Speech Hear. Res. **34**, 534–543.
Kreiman, J., Gerratt, B. R., and Berke, G. S. (**1994**). ''The multidimensional nature of pathologic vocal quality,'' J. Acoust. Soc. Am. **96**, 1291–1302.
Kreiman, J., Gerratt, B. R., Kempster, G. B., 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.
Kreiman, J., Gerratt, B. R., Precoda, K., and Berke, G. S. (**1992**). ''Individual differences in voice quality perception,'' J. Speech Hear. Res. **35**, 512–520.
Kuwabara, H., and Ohgushi, K. (**1984**). ''Experiments on voice qualities of vowels in males and females and correlation with acoustic features,'' Language Speech **27**, 135–145.
Laver, J., Hiller, S., and Beck, J. M. (**1992**). ''Acoustic waveform perturbations and voice disorders,'' J. Voice **6**, 115–126.
Lindau, M. (**1985**). ''The story of /r/,'' in *Phonetic Linguistics: Essays in Honor of Peter Ladefoged*, edited by V. Fromkin (Academic, Orlando), pp. 157–168.
Matsumoto, H., Hiki, S., Sone, T., and Nimura, T. (**1973**). ''Multidimensional representation of personal quality of vowels and its acoustical correlates,'' IEEE Trans. Audio Electroacoust. **AU-21**, 428–436.
Murry, T., and Singh, S. (**1980**). ''Multidimensional analysis of male and female voices,'' J. Acoust. Soc. Am. **68**, 1294–1300.
Murry, T., Singh, S., and Sargent, M. (**1977**). ''Multidimensional classification of abnormal voice qualities,'' J. Acoust. Soc. Am. **61**, 1630–1635.
Nieboer, G. L., De Graaf, T., and Schutte, H. K. (**1988**). ''Esophageal voice quality judgments by means of the semantic differential,'' J. Phon. **16**, 417–436.
SAS Institute Inc. (**1992**). SAS/STAT Software: Changes and Enhancements, Release 6.07 (SAS Tech. Rep. P-229) (SAS Institute, Cary, NC).
Schiffman, S., Reynolds, M., and Young, F. (**1981**). *Introduction to Multidimensional Scaling: Theory, Method, and Applications* (Academic, New York).
Takahashi, H., and Koike, Y. (**1975**). ''Some perceptual dimensions and acoustic correlates of pathological voices,'' Acta Otolaryngol. (Stockholm), Suppl. **338**, 2–24.
Titze, I. R. (**1992**). ''Acoustic interpretation of the voice range profile (phonetogram),'' J. Speech Hear. Res. **35**, 21–34.
Walden, B. E., Montgomery, A. A., Gibeily, G. J., Prosek, R. A., and Schwartz, D. M. (**1978**). ''Correlates of psychological dimensions in talker similarity,'' J. Speech Hear. Res. **21**, 265–275.
Wilson, F. B. (**1977**). *Voice Disorders* (Learning Concepts, Austin, TX).
Wirz, S., and Beck, J. M. (**1995**). ''Assessment of voice quality: The vocal profiles analysis scheme,'' in *Perceptual Approaches to Communication Disorders*, edited by S. Wirz (Whurr, London), pp. 39–55.
Wittgenstein, L. (**1958**). *The Blue and Brown Books. Preliminary Studies for the Philosophical Investigations* (Harper and Row, New York).
Wolfe, V., and Ratusnik, D. (**1988**). ''Acoustic and perceptual measurements of roughness influencing judgments of pitch,'' J. Speech Hear. Disord. **53**, 15–22.
Yumoto, E., Gould, W. J., and Baer, T. (**1982**). ''Harmonics-to-noise ratio as an index of the degree of hoarseness,'' J. Acoust. Soc. Am. **71**, 1544–1550.