Title
The multidimensional nature of pathologic vocal quality.

Journal
The Journal of the Acoustical Society of America, 96(3)

ISSN
0001-4966

Authors
Kreiman, J
Gerratt, B R
Berke, G S

Publication Date
1994-09-01

Peer reviewed

The multidimensional nature of pathologic vocal quality

Jody Kreiman, Bruce R. Gerratt, and Gerald S. Berke
Division of Head and Neck Surgery, UCLA School of Medicine, CHS 62-132, Los Angeles, California 90024 and Audiology and Speech Pathology (126), VA Medical Center, West Los Angeles, Wilshire & Sawtelle Boulevards, Los Angeles, California 90073

(Received 29 July 1992; revised 14 March 1994; accepted 5 May 1994)

Although the terms "breathy" and "rough" are frequently applied to pathological voices, widely accepted definitions are not available and the relationship between these qualities is not understood. To investigate these matters, expert listeners judged the dissimilarity of pathological voices with respect to breathiness and roughness. A second group of listeners rated the voices on unidimensional scales for the same qualities. Multidimensional scaling analyses suggested that breathiness and roughness are related, multidimensional constructs. Unidimensional ratings of both breathiness and roughness were necessary to describe patterns of similarity with respect to either quality. Listeners differed in the relative importance given to different aspects of voice quality, particularly when judging roughness. The presence of roughness in a voice did not appear to influence raters' judgments of breathiness; however, judgments of roughness were heavily influenced by the degree of breathiness, the particular nature of the influence varying from listener to listener. Differences in how listeners focus their attention on the different aspects of multidimensional perceptual qualities apparently are a significant source of interrater unreliability (noise) in voice quality ratings.

PACS numbers: 43.71.Bp, 43.71.Gv, 43.70.Dn

INTRODUCTION

"Breathy" and "rough" are among the most familiar labels for pathological voice qualities, and have been in common use since ancient times (see Laver, 1981, for review). Because of their importance for describing a wide variety of pathologies (e.g., Darley *et al.*, 1969; Isshiki *et al.*, 1969), these qualities are the subjects of frequent study in the literature on voice quality evaluation. For example, many papers have examined the correlation between acoustic and aerodynamic measures and rated breathiness and/or roughness (e.g., Arends *et al.*, 1990; Arnold and Emanuel, 1979; Coleman, 1969; Fritzell *et al.*, 1986).

However, despite a long history and extensive literature, the perceptual reality of breathiness, roughness, and related qualities (e.g., harshness, hoarseness) has never been systematically examined. In fact, these perceptual qualities have never received widely accepted definitions in the clinical literature, whether formal or informal. Thus it is difficult to determine precisely what a particular author means by "hoarseness," "harshness," "breathiness," "roughness," or any other label for vocal quality. The general lack of research into the perceptual reality and meaning of important descriptors of pathological voices is a long-standing problem in voice research (Jensen, 1965; Reed, 1980).

Further, listeners often disagree when they rate vocal qualities, suggesting that significant individual differences exist in the meaning assigned to such terms in practice. For example, Shipp and Huntington (1965) found interrater correlations (Pearson's r for pairs of raters) ranging from 0.33 to 0.78 for ratings of breathiness on an 8-point scale Kreiman *et al.* (1993) reported interrater correlations ranging from 0.55 to 0.92 for ratings of vocal roughness on a 7-point scale. Thus it appears that listeners may differ considerably from one another in the ratings assigned to any one voice, despite the fact that most individuals can rate voices consistently (see Kreiman *et al.*, 1993, for review). Understanding the sources of this listener variability in voice quality ratings might lead to the development of more reliable rating protocols.

Perhaps because of the lack of systematic research in this area, authors also disagree about the relationships among breathiness, roughness, and other vocal qualities. Two implicit views are prominent in the literature on pathological voices. In the first, different perceptual qualities are treated as independent features of voices which may reasonably be assessed individually. This view is implied by the many studies where ratings of a single quality are compared to objective measurements (e.g., Sansone and Emanuel, 1970; Wendahl, 1966; Yanagihara, 1967; Yumoto *et al.*, 1982; Yumoto *et al.*, 1984), and occasionally is assumed explicitly (e.g., Whitehead and Emanuel, 1974). In the second, breathiness and roughness are both treated as subordinate aspects of some other quality (a hierarchical view). For example, Fairbanks (1960) argued that breathiness and harshness are both components of a superordinate "hoarse" quality. (See also Laver, 1980, for discussion of a descriptive phonetic approach to similar qualities in normal phonation.)

Very little experimental evidence is available regarding either of the traditional views of vocal quality. The "independent feature" view is supported by studies finding that raters agree highly when they rate individual qualities (e.g., Klich, 1982; Lively and Emanuel, 1970; Sapir and Aronson, 1985). However, many other studies have reported low or variable levels of interrater reliability (e.g., Cullinan *et al.*, 1963; Nieboer *et al.*, 1988; Yumoto *et al.*, 1984; see Kreiman *et al.*, 1993, for review). Other studies provide limited support for a hierarchical view. Shipp and Huntington (1965) found ratings of breathiness and hoarseness were moderately

 0001-4966/94/96(3)/1291/12/$6.00

correlated for three of four expert raters, suggesting that these qualities are related, but not in any simple way. One factor analytic study also suggests that breathiness and roughness may be perceptually complex and interrelated: Hammarberg *et al.* (1980) found a "breathy-overtight" dimension (negatively associated with the scales "breathy," "wheezing," "lack of timbre," "moments of aphonia," "husky," and positively associated with "creaky/vocal fry") and a "coarse-light" dimension (positively associated with the scales "coarse," "rough," "harsh," and negatively associated with "high pitch," "middle register," and "restrained").

Other research has specifically addressed the perceptual structure of "hoarseness." Isshiki and Takeuchi (1970) used semantic differential techniques and factor analysis to examine subclassifications of hoarse voice quality. They found four factors, which they labeled "rough," "breathy," "asthenic" (lack of vocal strength), and "near-normal." The GRBAS protocol proposed by the Japanese Society of Logopedics (e.g., Hirano, 1981) maintained this distinction between rough, breathy, asthenic, and near-normal (the "grade" scale) aspects of hoarse voice, but added a scale for "strained" quality. Finally, Takahashi and Koike (1975) found that ratings of breathiness and roughness were moderately but significantly correlated ($r=0.47$), and concluded that the two qualities are not independent factors in a perceptual space. They also described factor analyses that supported Isshiki and Takeuchi's (1970) breathy and rough factors for the description of hoarseness.

Previous studies using multidimensional scaling (MDS) suggest that breathiness and roughness are important perceptual features of pathological voices, but that listeners differ from one another in how they judge these qualities. Kreiman *et al.* (1990) found dimensions correlated with rated breathiness and roughness in a MDS study of 18 pathological male voices. However, "rough" and "breathy" dimensions did not consistently emerge from a subsequent study examining individual differences in voice perception (Kreiman *et al.*, 1992), suggesting these dimensions are not perceptually important for every listener, even in a fixed perceptual context. Other MDS studies have not consistently produced breathiness and roughness dimensions. Murry *et al.* (1977) found dimensions associated with volume velocity (moderately correlated with breathiness ratings) and presence/absence of periodicity (related to rated hoarseness) in a study of pathological male voices.[1] In contrast, Kempster *et al.* (1991) found dimensions related to intensity, frequency, and perturbation in a MDS study of dysphonic female speakers. They did not speculate as to how these dimensions might relate to traditional labels for voice quality.

Thus a number of issues remain unresolved, both with respect to the perceptual status of breathiness and roughness and to the perception of pathological voices in general. In particular, the relationship among different labels for vocal quality has never been systematically investigated. Thus it is unclear whether listeners can rate different (unidimensional) voice qualities independently, or whether qualities are better viewed as multidimensional constructs whose dimensions may influence one another during the rating process. Such information is essential for designing valid and reliable protocols for clinically evaluating pathological voice quality.

The present study combined multidimensional scaling (MDS) and unidimensional rating approaches to address these issues directly. MDS techniques have several advantages over more commonly used rating methods. They do not require *a priori* assumptions about the dimensionality of a quality, and thus allow unbiased investigation into the number of dimensions necessary to explain listeners' judgments, the nature of such dimensions, and the relationships among them. They also permit detailed examination of differences among listeners in the criteria used to rate a voice on some quality scale. The addition of unidimensional ratings of the same voices allowed us to relate multidimensional results to traditional impressionistic labels for voice quality, to determine how listeners may map multidimensional qualities onto unidimensional rating scales.

I. PERCEPTUAL SPACES FOR BREATHINESS AND ROUGHNESS

A. Method

1. Listeners (group 1)

Five native speakers of English participated in this study. None had participated in previous studies using these stimuli. Two were speech pathologists (listeners 1 and 2), two were linguists specializing in voice research (listeners 3 and 4), and one was trained in both linguistics and speech pathology (listener 5). All were trained in the American tradition of voice quality description, and each had at least three years postgraduate experience judging voices. Listeners worked with pathological voices on a daily basis, and regularly applied the terms studied here. Listeners reported no history of voice, speech, language, or hearing difficulties.

2. Stimuli

The voices of 18 male speakers with voice disorders were selected at random from a library of audio recordings made as part of a phonatory function evaluation. During this evaluation, speakers sustained the vowel /ɑ/ at conversational levels of pitch and loudness. Speakers varied widely in the overall severity of their voice disorder. Mildly, moderately, and severely breathy and rough voices were all represented, as were a variety of diagnoses. A previous multidimensional scaling study using these voices (Kreiman *et al.*, 1990) revealed breathiness and roughness dimensions which each accounted for more than 25% of the variance in listeners' dissimilarity judgments.[2]

Voice samples were low-pass filtered using two 4-pole Butterworth filters with cutoff frequencies of 6300 Hz, and two with cutoff frequencies of 7500 Hz, for a total reduction in amplitude of 3.2 dB at 5.6 kHz and 39.4 dB at 9 kHz. They were then sampled at 17.8k samples/second using a 16-bit A/D converter. A 1.7-s sample was taken from the middle portion of each speaker's /ɑ/. The digitized segments were normalized for peak voltage, and onsets and offsets were multiplied by 10-ms ramps to eliminate click artifacts. Stimuli were then output through a 16-bit D/A converter using the same filter settings.

Two experimental tapes were constructed. Each included both orders of all possible pairs of the 18 pathological voices (excluding pairs where voices were the same), for a total of 306 trials per tape. Stimuli were rerandomized for the second tape. For both tapes, voice samples within a pair were separated by 1 s, and pairs were separated by 6 s.

3. Procedure

Each listener participated in two listening sessions separated by at least one week. Testing took place in a sound-treated booth. At one session listeners judged the dissimilarity of each pair of voices with respect to levels of breathiness; at the other they judged dissimilarity with respect to roughness. One experimental tape was used for the first session, and the other at the second, so each listener made two judgments of each quality for each pair of voices. Order of task and tape presentation was randomized across listeners.

Listeners rated the dissimilarity of the pairs of voices on 7-point equal-appearing interval scales, where "1" represented identical levels of breathiness/roughness and "7" represented extreme difference in breathiness/roughness levels. Thus a rating of "1" could mean voices were both very breathy, not breathy at all, and so on, while a rating of "7" meant that one voice was (near-) normal and one was severely breathy or rough. Formal definitions of breathiness and roughness were not offered. Instead, listeners were asked to use whatever standards they normally applied in their clinical practice or research. They were instructed to focus their attention on the quality being judged and to ignore any other qualities the voices might have. They were also asked to judge each pair of voices as independently as possible, and were discouraged from changing previous responses after hearing a new pair of voices.

Each test session lasted approximately 1.5 h. Listeners were encouraged to take brief breaks during this period as needed.

4. Multidimensional scaling analyses

Previous studies indicate that presentation order has a significant effect on listener judgments of vocal quality, because the first member of a pair of voices provides a context against which the second is judged, highlighting different facets of these complex stimuli (Gerratt *et al.*, 1993). To avoid losing such information, matrices of dissimilarity judgments were not symmetrized across the diagonal. Instead, each listener's judgments for each task were assembled into two half-matrices (upper and lower halves, minus the diagonal). Judgments from all listeners were combined and analyzed using the nonmetric individual differences model (INDSCAL) of SAS PROC MDS (Kruskal and Wish, 1978; SAS Institute, 1992; Schiffman *et al.*, 1981). Separate analyses were undertaken for breathiness and roughness judgments. Each analysis included ten half-matrices (two from each listener).

Scaling solutions were found in one to six dimensions for each rating task. Based on values of stress, on the amount of variance accounted for by each solution (R^2; Table I), and on interpretability, two-dimensional solutions were selected for both the breathiness and roughness judgments.

TABLE I. R^2 (variance accounted for) and stress for the group multidimensional scaling solutions.

No. of dimensions in solution	Rating task: Breathiness judgments		Rating task: Roughness judgments	
	R^2	Stress	R^2	Stress
6	0.75	0.17	0.81	0.19
5	0.73	0.19	0.81	0.20
4	0.71	0.23	0.80	0.23
3	0.68	0.27	0.76	0.26
2	0.67	0.33	0.75	0.30
1	0.62	0.42	0.70	0.36

5. Acoustic analyses

To assist in interpreting the scaling solutions, a number of time- and frequency-domain measurements were made on the test voices.[3] The fundamental frequency ($F0$) and the frequencies of the first three formants ($F1$, $F2$, and $F3$) were measured from spectrographic displays. $F0$ was measured from narrow-band displays with a frequency range of 0–1 kHz; the center frequencies of the three clearest harmonics were measured to ensure accuracy. Formants were measured with reference to both narrow- and wideband displays (with a frequency range of 0–4 kHz), and to displays of line spectra of the vowels. Measurements were taken from sections of the display where the formants appeared most steady and level.

For jitter and shimmer measurements, a point on each waveform cycle that could be identified reliably from cycle to cycle was selected interactively. Measurements of mean jitter, standard deviation of jitter, and the coefficient of variation for jitter were then calculated using parabolic interpolation when the point marked was a peak and linear interpolation when a zero crossing was marked (Titze *et al.*, 1987). Analogous shimmer measurements were also calculated, using the difference in dB between the highest and lowest points in each marked cycle as the amplitude.

Several additional acoustic measures were also obtained. The natural logarithm of the standard deviation of the period lengths (LNSD; see Wolfe and Steinfatt, 1987) was calculated for each voice sample, as were the harmonics-to-noise ratio (HNR) (Yumoto *et al.*, 1982) and the ratio of the amplitudes of the first to the second harmonic (H1–H2) (Bickley, 1982; Ladefoged, 1981). Finally, we calculated the "partial period comparison" (PPC), a time-domain comparison of the standard deviations of differences between moving vectors, each about 0.6 times the estimated period length (Ladefoged *et al.*, 1988). To generalize this measure to long segments of speech, it was applied to a sample approximately three glottal cycles long; the next two cycles were skipped, the next three measured, continuing in this manner for the duration of the vowel sample. The mean of the indices generated for the entire voice sample was then calculated for use in this study. The PPC is moderately correlated with mea-

sures of jitter and shimmer, and may measure variability within an utterance in levels of signal unsteadiness.

6. Unidimensional perceptual measures of vocal quality

To assess the extent to which multidimensional spaces capture the information available from traditional unidimensional ratings of voice qualities, we gathered additional ratings of the stimulus voices using equal-appearing interval (EAI) scales for breathiness and roughness. A second group of eight expert listeners (four speech pathologists and four otolaryngologists) participated in this experiment. As above, listeners were trained in the American tradition of voice quality description, and each had a minimum of 3 years experience evaluating pathological voices. None had previous experience with these stimulus voices, and none participated in the study described above.

Judgments of breathiness and roughness were made at separate test sessions at least one week apart. Voices were rerandomized for each listener and rating task. Testing took place in a sound-treated booth. Stimulus digitization and playback were as described above. The rate of stimulus presentation was controlled by the listener.

At each listening session, listeners rated the voices using 7-point EAI scales. On these scales the value "1" represented minimum breathiness or roughness, and "7" represented severe breathiness or roughness. Scale endpoints were labeled accordingly. Listeners were asked to pay attention only to the quality being judged, and to ignore any other qualities the voice might have. No formal definitions of breathiness or roughness were offered. Instead, listeners were asked to use whatever standards they usually applied in their clinical practices. Listeners were able to replay the voices if necessary before making their judgments.

B. Results

1. Independence of breathiness and roughness ratings

To determine if dissimilarity judgments of breathiness and roughness were independent, we examined the correlation between the two sets of unscaled ratings for each of the five listeners in group 1. Values of Pearson's r averaged 0.25 (standard deviation=0.10). Unidimensional (EAI) ratings of breathiness and roughness for group 2 also were not highly correlated (mean Pearson's r=0.27, s.d.=0.17). These values are so low as to suggest that judgments of the two qualities were in fact independent, for both the dissimilarity ratings and the EAI task.

2. Rating reliability

Intrarater (test–retest) reliability for dissimilarity ratings of breathiness and roughness was assessed by calculating the correlation (Pearson's r) between the first and second rating of each pair of voices. Values for breathiness ratings ranged from 0.34 to 0.68, with a mean of 0.55. However, across listeners 72.5% of repeated ratings differed by one scale value or less (range=60.8%–90%; chance=38.8%), suggesting the low correlation values reflect the limited range of the rating scale. Similarly for the roughness ratings, repeated dissimilarity ratings were not particularly well correlated (mean r=0.62; range=0.37–0.81), but the majority of repeated ratings were within one scale value (73.3%; range =58.8%–85.6%). Note that these values also reflect the effects of the different presentation orders used for the first and second ratings of the voice pairs.

For the EAI ratings (listener group 2), values of Pearson's r comparing a listener's first and second ratings of breathiness ranged from 0.63 to 0.93, with a mean of 0.81 (s.d.=0.11). For roughness, intrarater correlations ranged from 0.66 to 0.91, with a mean of 0.78 (s.d.=0.08). For interrater reliability, values of Pearson's r comparing all possible pairs of listeners indicated that individual listeners did not necessarily agree particularly well with one another. For breathiness, mean Pearson's r for pairs of raters was 0.69 (s.d.=0.11, range=0.44–0.86). For roughness, mean Pearson's r=0.54 (s.d.=0.20, range=0.05–0.82). One listener in particular rated vocal roughness consistently (intrarater Pearson's r=0.77), but differed considerably from the rest of the group (average r=0.35; range=0.05–0.62). When that listener was excluded, the mean correlation among raters for roughness ratings increased to 0.60 (s.d.=0.17, range=0.22–0.82).

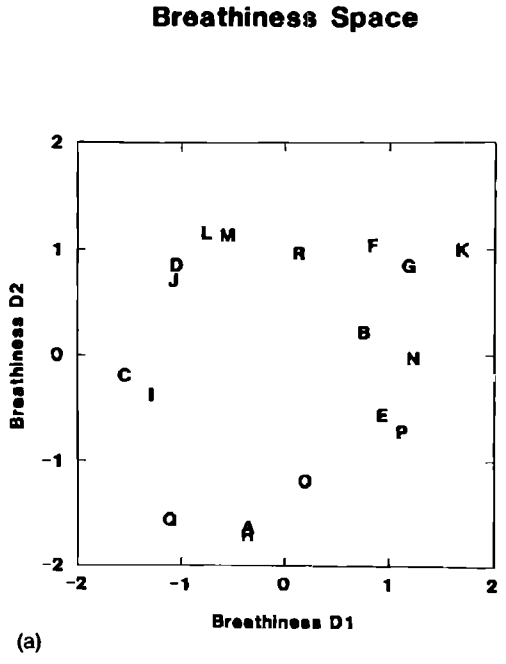
However, average EAI ratings were sufficiently reliable for our purposes (e.g., Berk, 1979). For breathiness ratings, the intraclass correlation=0.93 [model (2,8); e.g., Ebel, 1951; Shrout and Fleiss, 1979]; for roughness, ICC(2,8) =0.86. Accordingly, average EAI ratings were used for interpreting group perceptual spaces. We will discuss issues surrounding the reliability of individual raters in more detail in Sec. III below.

3. Multidimensional scaling solutions

As described above, two-dimensional solutions were selected for both the breathiness and roughness data. Stimuli were arranged roughly in a circle in both spaces (Fig. 1). The breathiness space accounted for 67% of the variance in the underlying dissimilarity ratings of breathiness; the first dimension (D1) accounted for 52% of the variance, and the second dimension (D2) for 15%.[4] The roughness solution accounted for 75% of the variance in the underlying dissimilarity ratings, with D1 contributing 55% and D2 contributing 20% to the explained variance.

Scaling solutions were interpreted by examining significant correlations between stimulus coordinates on each dimension and the acoustic parameters described above. In cases where dimensions were significantly correlated with more than one acoustic variable, multiple regression was used to determine whether the acoustic variables accounted for independent aspects of variance in stimulus coordinates.

Interpretations for the two spaces were similar. D1 in the breathiness space was correlated with a weighted sum of H1–H2 and LNSD (R=0.84); D2 was correlated with a weighted sum of mean shimmer and $F0$ (R=0.82). In the roughness space, stimuli on D1 were ordered according to a weighted sum of mean shimmer and H1–H2 (R=0.91). D2 correlated highly with a weighted sum of H1–H2 and $F0$ (R=0.81).



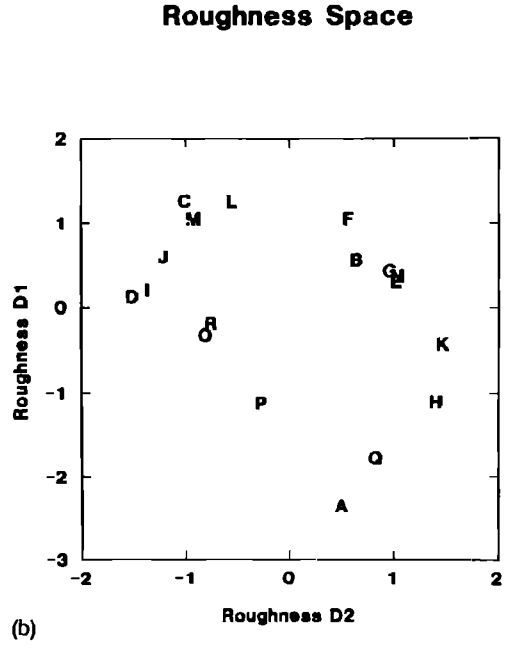


FIG. 1. Multidimensional scaling solutions for the combined subject group. Letters indicate individual voices. A: Space derived from dissimilarity ratings of breathiness. B: Space derived from dissimilarity ratings of roughness.

Stimulus coordinates in the two spaces were significantly correlated. Breathiness D2 corresponded to roughness D1 ($r=0.74$, $p<0.01$), and breathiness D1 corresponded to roughness D2 ($r=0.66$, $p<0.01$). These findings suggest that similar, multidimensional perceptual structures underlie ratings of vocal breathiness and roughness for our listeners.

4. Unidimensional versus multidimensional ratings

To examine how adequately unidimensional ratings capture this multidimensional information, multiple regression was used to compare stimulus coordinates on the dimensions derived above to average EAI ratings of breathiness and roughness for the same voices. Results are given in Table II.

TABLE II. Results of multiple regressions comparing EAI ratings and stimulus coordinates in the group perceptual spaces.

Dimension	Standardized regression coefficients: Breathiness ratings	Standardized regression coefficients: Roughness ratings	$F(2,15)$[a]	Multiple R^2
Breathy space D1	1.03	−0.23	88.99	0.92
Breathy space D2	···	0.84	25.53	0.77
Rough space D1	−0.26	0.99	39.87	0.84
Rough space D2	−0.74	0.58	8.26	0.52

[a]All F values are significant at $p<0.01$.

Unidimensional ratings did capture the majority of the information in both of the multidimensional spaces. Values of multiple R^2 ranged from 0.52 to 0.92 for the four dimensions. D1 in the breathiness space was significantly related to unidimensional ratings of both breathiness and roughness; D2 in the breathiness space was related only to rated roughness. Similarly, both dimensions in the roughness space were significantly related to both sets of unidimensional ratings. D1 corresponded primarily to rated roughness, while D2 was strongly related to both breathiness and roughness. Thus EAI ratings of both breathiness and roughness were needed to describe each single perceptual space.

5. Individual differences in perceptual strategy

The R^2 values, squared subject weights, and weirdness values for individual subjects in the multidimensional scaling analyses are given in Table III. Group values are included for comparison. The R^2 values in Table III represent the amount of variance in that subject's ratings that is accounted for by the group scaling solution. That is, they measure the overall fit of the group solution to an individual's data. Subject weights reflect the perceptual importance a given dimension has for an individual subject. The weights printed by the MDS procedure have been squared, and sum to R^2 for a given subject. Weirdness reflects the extent to which an individual's weights on the dimensions are proportional to the group weights. A subject with weights proportional to the average weights has a weirdness of 0; a subject with one very large weight and one small weight has a weirdness near 1.

As Table III shows, subjects differed considerably in the extent to which the group solution reflected their perceptual strategies. For the breathiness space, weirdness values suggest that listeners 3 and 5 differed substantially (and in different directions) from average in the relative importance given the two dimensions. Listener 3 relied relatively heavily on D1 (which was correlated with a weighted sum of H1–H2 and LNSD). Because this dimension was much more important overall, this subject's R^2 value is high relative to the total group. In contrast, dimension 2 (correlated with a weighted sum of shimmer and $F0$) was much more important for listener 5 than for the group as a whole. Consequently, R^2 for this listener is lower than that for the group.

TABLE III. R^2 values, squared subject weights, and weirdness values for individual listeners in the group scaling solutions. Note: Each value represents the average of scores for the top and bottom half matrix for that listener. See text for discussion.

	Breathiness space			Roughness space		
Listener/Matrix	R^2	Squared subject weights (D1/D2)	Weirdness	R^2	Squared subject weights (D1/D2)	Weirdness
1/1	0.56	0.42/0.14	0.07	0.59	0.29/0.30	0.34
1/2	0.63	0.56/0.07	0.26	0.68	0.38/0.30	0.27
2/1	0.66	0.58/0.08	0.21	0.71	0.53/0.18	0.00
2/2	0.72	0.61/0.11	0.13	0.72	0.59/0.13	0.14
3/1	0.81	0.74/0.07	0.31	0.80	0.52/0.28	0.16
3/2	0.75	0.68/0.07	0.29	0.85	0.45/0.40	0.31
4/1	0.78	0.59/0.19	0.05	0.78	0.58/0.20	0.02
4/2	0.76	0.63/0.13	0.08	0.83	0.72/0.11	0.24
5/1	0.61	0.26/0.35	0.47	0.80	0.78/0.02	0.66
5/2	0.45	0.18/0.27	0.49	0.71	0.64/0.07	0.34
Group	0.67	0.53/0.16		0.75	0.55/0.20	

Listeners 1, 3, and 5 differed from the average in their dimension weights for the roughness space. D2 (interpreted as a weighted sum of H1–H2 and $F0$) was more important for listeners 1 and 3 than for the group as a whole. D1 (interpreted as a weighted sum of shimmer and H1–H2) was less important for listener 1 than for the group, and more important for listeners 4 and 5. R^2 values are somewhat lower only for listener 1, who differed from the group in the relative perceptual salience of both dimensions.

C. Discussion

These data suggest that breathiness and roughness are not independent unidimensional aspects of vocal quality. Rather, they appear to be different aspects of a single multidimensional "quality." Rated breathiness and roughness were both needed to describe each of the MDS spaces. Thus it appears that having listeners rate only breathiness or roughness is not adequate to assess the extent to which a voice possesses that individual quality. Our results indicate that information about both qualities is needed to measure either.

These data further suggest that consistent perceptual differences may underlie stable group scaling solutions. In particular, listener 3 appears to have relied more heavily on dimensions correlated with H1–H2 than did the average listener; and listener 5 apparently judged both breathiness and roughness in terms of acoustic signal perturbation.

To examine listener differences in more detail, we undertook a second set of multidimensional scaling analyses. Each new analysis included data from a single subject for a single vocal quality. We hoped such analyses would provide insight into the nature and extent of intersubject variability in perception of breathiness and roughness.

II. SCALING SOLUTIONS FOR INDIVIDUAL LISTENERS

A. Method

Separate breathiness and roughness spaces were calculated for each individual listener in group 1 using the INDSCAL model of SAS PROC MDS, as above. Each analysis included the top and bottom half matrices of dissimilarity judgments produced by that listener. This procedure was

TABLE IV. Interpretations for individual perceptual spaces: breathiness ratings.

Subject	R^2 for solution	Dimension	Weight	Interpretation	R for interpretation
1	0.75	1	0.63	H1–H2+HNR	0.83
		2	0.11	shimmer SD	0.59
2	0.84	1	0.74	H1–H2+LNSD+PPC	0.89
		2	0.10	shimmer coeff. of var.+jitter SD	0.73
3	0.86	1	0.69	shimmer coeff. of var.	0.62
		2	0.16	$F0$+H1–H2	0.71
4	0.87	1	0.51	PPC+H1–H2	0.78
		2	0.36	H1–H2+LNSD	0.88
5	0.66	1	0.39	shimmer coeff. of var.+$F0$	0.70
		2	0.27	HNR+H1–H2	0.76

TABLE V. Interpretations for individual perceptual spaces: roughness ratings.

Subject	R^2 for solution	Dimension	Weight	Interpretation	R for interpretation
1	0.70	1	0.42	mean shimmer+H1–H2	0.87
		2	0.28	$F0$+H1–H2+$F3$	0.91
2	0.82	1	0.72	PPC+H1–H2	0.93
		2	0.10	H1–H2	0.58
3	0.94	1	0.68	PPC+H1–H2	0.90
		2	0.26	$F0$	0.62
4	0.94	1	0.76	$F0$+H1–H2+PPC	0.86
		2	0.18	jitter SD	0.75
5	0.86	1	0.53	PPC	0.50
		2	0.33	PPC+$F0$	0.81

chosen over the traditional practice of symmetrizing matrices by averaging data points across the diagonal, because it preserves context-dependent information about vocal qualities, as argued above. Analyses of symmetrized data confirmed that one-dimensional solutions were appropriate for the symmetrized data, while solutions for the unsymmetrized data provided both higher overall R^2 values and more reasonable interpretations. Solutions were calculated in one to four dimensions. Based on values of stress, R^2, and interpretability, two-dimensional solutions were selected for all ten analyses.

B. Results

1. Variance accounted for by the scaling solutions

R^2 values for the individual scaling solutions are given in Tables IV and V. In every case, scaling solutions accounted for the majority of the variance in an individual's dissimilarity ratings. For the breathiness spaces, R^2 values ranged from 0.66 to 0.87, with a mean of 0.80 (s.d.=0.09). R^2 values for the roughness spaces ranged from 0.82 to 0.94, with a mean of 0.85 (s.d.=0.10).

2. Interpretation of the individual spaces

Individual perceptual spaces were interpreted using the methods described above (Sec. I B 3). Results for the breathiness spaces are included in Table IV, and for the roughness spaces in Table V. Stimulus configurations for individual breathiness spaces are shown in Fig. 2, and for roughness spaces in Fig. 3.

For breathiness, both dimension interpretations and stimulus configurations suggest that listeners are differentially weighing a fairly constant set of acoustic cues. There is little evidence of gross differences in perceptual strategy. Each listener's first dimension (which accounts for the majority of variance in the solutions) is significantly correlated with at least one dimension in the space for every other listener (r=0.64–0.93). Across solutions only one dimension (D2 for listener 2) was not significantly related to any dimension in any other space.[5] All five spaces are interpretable in terms of H1–H2 and perturbation; and all contain two fairly continuous dimensions, with no obvious clusters of stimuli.

Listeners do differ in the relative weight given each dimension. For example, listeners 3 and 5 relied more on shimmer and $F0$, and less on H1–H2, than did listeners 1, 2, and 4; and listeners 1 and 5 attended to spectral noise (as measured by the harmonics-to-noise ratio) in addition to spectral slope (as measured by H1–H2). Overall, however, listeners seem to have used a fairly consistent set of perceptual parameters when judging the relative breathiness of the stimuli. This finding is consistent with the relatively high levels of interrater reliability for EAI ratings of breathiness reported above for these voices (Sec. I B 2).

Listeners differed more in their roughness judgments. As Table V shows, only listener 1's scaling solution matched the group solution. Listener 2 apparently did not use $F0$ when judging roughness, while listener 3 relied much more heavily on $F0$ than the average listener. Listener 4 also relied heavily on $F0$, but also apparently referred to levels of vocal jitter when judging roughness. Finally, neither dimension in the space for listener 5 was significantly correlated with H1–H2, although that parameter was important for the other four listeners.

Figure 3 further suggests that some listeners judged roughness using continuous scales, while others seem to have used features that are dichotomous or trichotomous. The perceptual spaces for listeners 1 and 2 suggest these subjects judged roughness using two features that varied continuously. However, listener 3's perceptual space shows three tight clusters of stimuli. The first dimension divides the stimuli into those with fairly steady phonation and those with a tremulous or unsteady quality. The second dimension is well correlated with $F0$; however, the upper cluster in this space includes voices which are simultaneously rough and breathy, while those in the lower cluster lack salient breathiness. Thus the space for listener 3 reflects a ternary division of voices into tremulous, breathy-rough, and other qualities. In contrast, the space for listener 4 is divided along the major diagonal, with tremulous voices in quadrant IV and others in quadrant II. The space for listener 5 also includes a cluster of tremulous voices, along with a small group of voices which seem to share a vowel quality nearer /æ/ than /ɑ/, and a cluster including the remainder of the voice set. Listeners 4 and 5 apparently treated roughness as a binary feature of voices (±tremulous), as compared to the ternary tremulous/breathy–rough/other distinction used by listener 3.

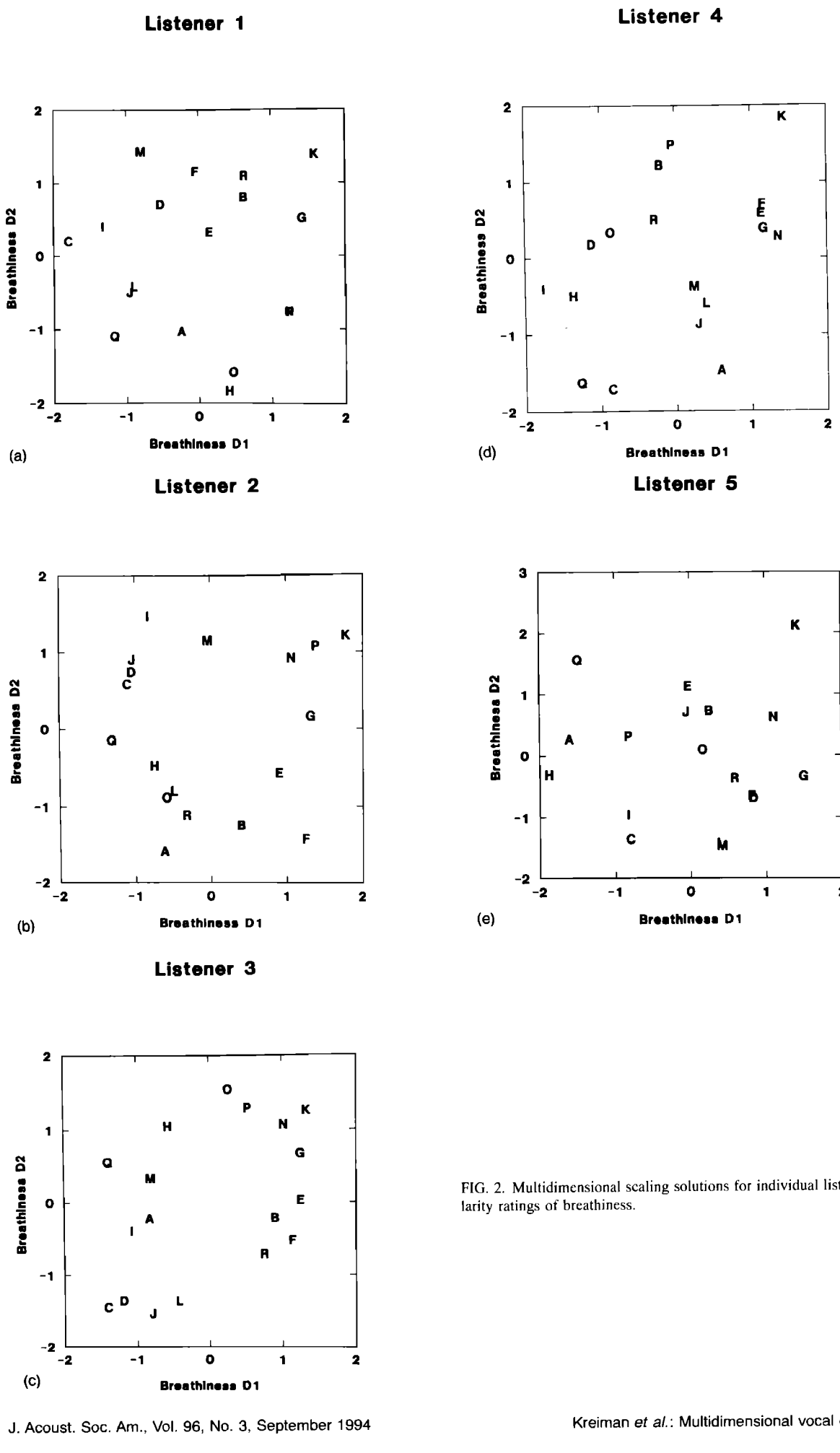


FIG. 2. Multidimensional scaling solutions for individual listeners' dissimilarity ratings of breathiness.

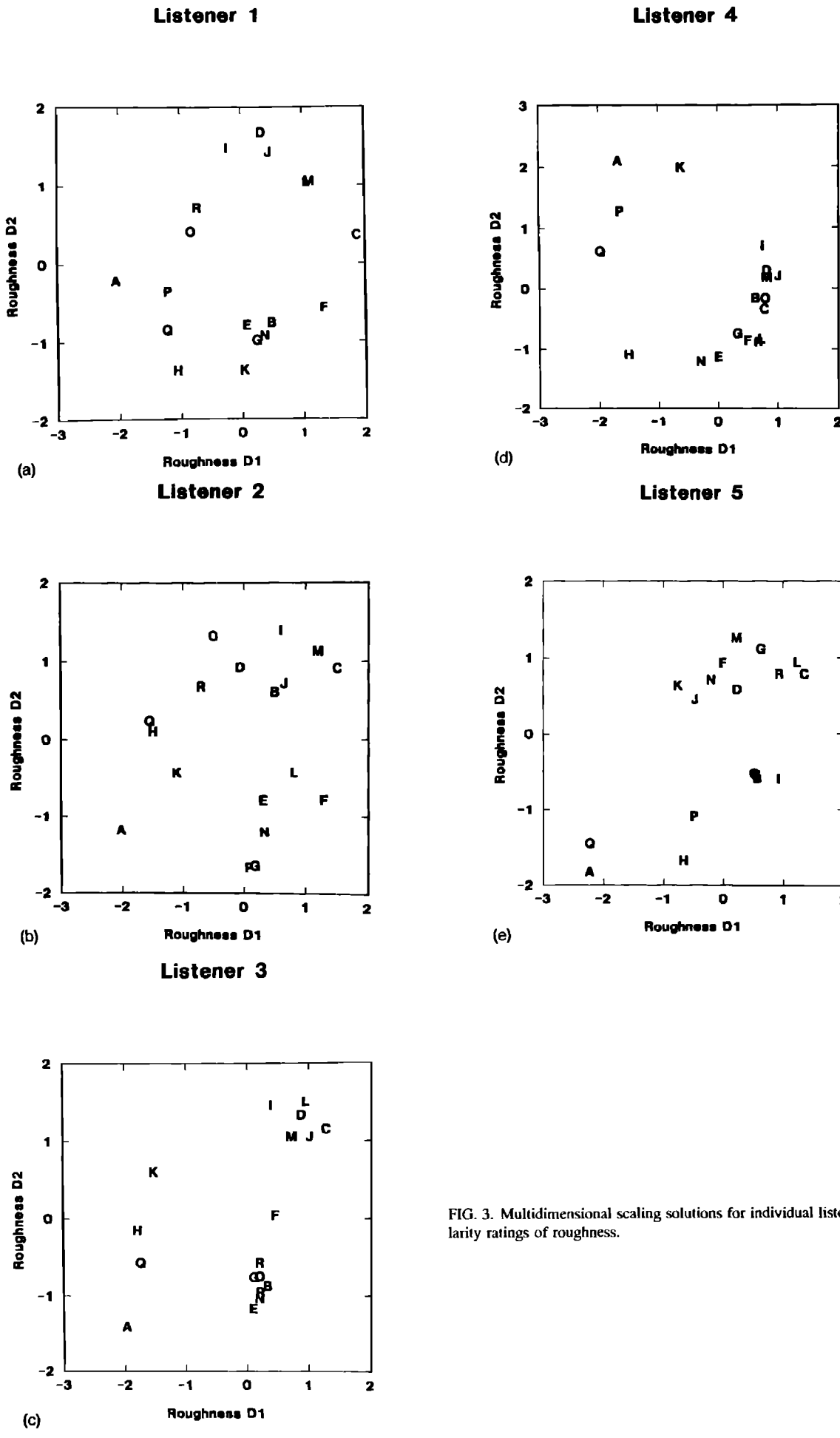


FIG. 3. Multidimensional scaling solutions for individual listeners' dissimilarity ratings of roughness.

TABLE VI. Significant correlations between EAI ratings and perceptual dimensions in the group MDS spaces.

Listener	EAI breathiness ratings correlated with:	EAI roughness ratings correlated with:
1	breathiness: D1 ($r=-0.78$)	roughness: D1 ($r=0.88$)
	roughness: ···	breathiness: D2 ($r=0.77$)
2	breathiness: D1 ($r=0.80$)	roughness: ···
	roughness: ···	breathiness: D2 ($r=0.75$)
3	breathiness: D1 ($r=0.80$)	roughness: ···
	roughness: ···	breathiness: ···
4	breathiness: D1 ($r=0.81$)	roughness: D1 ($r=0.73$), D2 ($r=0.67$)
	roughness: ···	breathiness: ···
5	breathiness: ···	roughness: ···
	roughness: ···	breathiness: D2 ($r=0.81$)
6	breathiness: D1 ($r=0.94$)	roughness: D1 ($r=0.80$)
	roughness: ···	breathiness: ···
7	breathiness: D1 ($r=0.82$)	roughness: D1 ($r=0.92$)
	roughness: ···	breathiness: D2 ($r=0.69$)
8	breathiness: D1 ($r=0.84$)	roughness: D1 ($r=0.74$)
	roughness: ···	breathiness: D2 ($r=0.83$)

III. PERCEPTUAL STRATEGIES AND INTERRATER RELIABILITY

As mentioned above, many researchers have reported low or variable levels of interrater reliability in studies of vocal quality, and reliability remains a serious problem in designing and using vocal quality rating systems in the clinic. The present findings suggest that one source of rating unreliability may be the tendency of different listeners to focus selectively on one or the other dimension of a given vocal quality. To test this hypothesis, we examined the correlations between the unidimensional (EAI) ratings of breathiness and roughness for listeners in group 2, and the coordinates of the voice stimuli in the group multidimensional spaces. Significant correlations ($p<.01$, adjusted for multiple comparisons) are listed in Table VI.

This table suggests that listeners consistently focused their attention on a single dimension of breathiness when making EAI ratings. With the exception of listener 5, whose perceptual strategy apparently deviated substantially from that of other listeners, EAI ratings were well correlated with D1 in the breathiness space, and were not correlated with any dimensions in the roughness space. This finding is consistent with the relatively high levels of interrater reliability reported above for breathiness ratings.

In contrast, listeners differed considerably in the way they focused their attention while judging vocal roughness. As Table VI shows, across listeners EAI ratings were correlated with a variety of perceptual dimensions, suggesting listeners varied considerably in the perceptual strategy they applied to the EAI task. Recall that D1 in the roughness space and D2 in the breathiness space were both correlated primarily with vocal shimmer; roughness D1 was also correlated with H1–H2, and breathiness D2 with $F0$. Thus listeners whose EAI ratings are correlated with D1 in the roughness space apparently attended to simultaneous breathiness when judging roughness; those whose ratings correlated primarily with D2 in the breathiness space attended to $F0$ rather than to breathiness.

Interrater reliability for the roughness ratings varied significantly with apparent perceptual strategy. EAI ratings for listeners who apparently shared no inferred perceptual features (e.g., listeners 1 and 3; listeners 2 and 4) were poorly correlated (average Pearson's $r=0.35$; s.d.$=0.15$). Ratings for listeners whose inferred perceptual strategies had one feature in common (e.g., listeners 1 and 4; listeners 2 and 7) were better correlated (mean Pearson's $r=0.64$; s.d.$=0.12$). Listeners whose EAI ratings were correlated with two dimensions in the perceptual spaces (e.g., listeners 1 and 7, listeners 7 and 8) had the highest levels of interrater reliability (mean Pearson's $r=0.73$, s.d.$=0.08$). A one-way ANOVA comparing Pearson's r values for these three groups of subjects showed a significant effect of inferred perceptual strategy on interrater agreement [$F(2,25)=17.98$, $p<0.01$]. Scheffé *post hoc* comparisons showed that listeners who shared one or two perceptual features agreed significantly better than those who shared no features. No difference was found between pairs of listeners sharing one or two features, possibly because of the small number of listeners in the latter group. These findings suggest that differences among listeners in how they focus attention on the different facets of vocal roughness is a significant cause of interrater unreliability.

IV. GENERAL DISCUSSION

The limits of multidimensional scaling spaces as perceptual models are well known and have been discussed elsewhere (e.g., Yost, 1989). Further, the present study used a relatively small set of male voices and a limited number of listeners. The perceptual features of female voices may differ significantly from those found for males, and interactions between listeners and speaker sex may occur (e.g., Batstone and Tuomi, 1981).

Nevertheless, our findings suggest that breathiness and roughness are related, multidimensional constructs. Most of the multidimensional information available from dissimilarity ratings was captured by the two sets of EAI ratings; however, EAI ratings of breathiness and roughness were both necessary to describe patterns of similarity with respect to either quality. Listeners differed in the relative importance given to different aspects of vocal quality, particularly when judging roughness. Simultaneous roughness did not appear to influence raters' judgments of breathiness; however, judgments of roughness were heavily influenced by degree of breathiness, the particular nature of the influence varying from listener to listener. Differential attention to different aspects of a quality is apparently a significant source of interrater unreliability in ratings of pathological voices.

This study indicates how traditional rating methods and scales may incorporate unsuspected sources of error. Problems of reliability have long plagued ratings of voice quality. We have previously argued that a significant portion of this unreliability in fact represents regular, predictable variability due to context effects, differences among listeners in background and perceptual strategy, characteristics of the task used to gather ratings, and interactions among these factors (Kreiman *et al.*, 1993). The present results confirm the importance of differences among listeners in modeling voice

perception. These differences range from dramatic (e.g., using unrelated perceptual strategies) to subtle (e.g., using continuous versus categorical dimensions for similar "features"). Listeners who differ more from one another agree less in their ratings; and listeners who differ less agree better.

Thus it appears that the multidimensional nature of the acoustic voice signal greatly influences unidimensional ratings of voice quality. Our results strongly suggest that a given vocal quality cannot be evaluated reliably out of the context of other qualities a voice may possess. However, it may be possible to develop voice rating protocols that control this source of variability. A recent study (Gerratt *et al.*, 1993) used an "anchored" EAI scale for vocal roughness, where each scale point was explicitly represented by a voice demonstrating that magnitude of vocal roughness. By fixing listener attention on a single dimension of roughness, this protocol produced significant improvements in interrater reliability relative to unanchored ratings. Such protocols may increase the likelihood that listeners will use similar perceptual strategies when judging a particular dimension.

The present study also highlights the need for more extensive, systematic investigation of the perceptual attributes of pathological voices and of the relationships among traditional terms for voice. The issues of what qualities are perceptually real and perceptually independent have been too long ignored in a field that is founded largely on perception (e.g., Jensen, 1965; Kreiman *et al.*, 1993). Rating protocols that reflect the natural perceptual categories inherent in the population to be rated should be easier to use, more valid, and more reliable than those using arbitrary labels and categories whose meaning is questionable. Increased attention to these matters will benefit both research and clinical practice.

ACKNOWLEDGMENTS

We thank Norma Antonanzas-Barroso for programming support, and Andrew Erman for his help in preparing the stimulus tapes. This research was supported in part by NIDCD Grant No. DC 01797 and by VA Merit Review Funds.

[1]Three other dimensions were reported: ± tumor, *F*0, and one uninterpreted dimension.

[2]The only other dimension to emerge from this analysis accounted for 23% of the variance in dissimilarity judgments, and was significantly correlated with *F*0 for the voices.

[3]The voice of one pathological speaker was clearly diplophonic. Thus only formant measurements were available for him.

[4]The variance accounted for by each dimension is reported by the scaling program.

[5]Although this dimension shares acoustic correlates with D1 for listeners 3 and 5, stimuli are arranged differently. Listener 2 emphasized differences among mildly-to-moderately pathological voices and compressed those between severely disordered voices, leading to low correlations with other listeners' perceptions.

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

Arnold, K. S., and Emanuel, F. **(1979)**. "Spectral noise levels and roughness severity ratings for vowels produced by male children," J. Speech Hear. Res. **22**, 613–626.

Batstone, S., and Tuomi, S. K. **(1981)**. "Perceptual characteristics of female voices," Lang. Speech **24**, 111–123.

Berk, R. **(1979)**. "Generalizability of behavioral observations: A clarification of interobserver agreement and interobserver reliability," Am. J. Mental Deficiency **83**, 460–472.

Bickley, C. **(1982)**. "Acoustic analysis and perception of breathy vowels," MIT., R.L.E. Speech Communications Group: Working Papers **1**, 71–82.

Coleman, R. F. **(1969)**. "Effect of median frequency levels upon the roughness of jittered stimuli," J. Speech Hear. Res. **12**, 330–336.

Cullinan, W. L., Prather, E. M., and Williams, D. E. **(1963)**. "Comparison of procedures for scaling severity of stuttering," J. Speech Hear. Res. **6**, 187–194.

Darley, F., Aronson, A., and Brown, J. **(1969)**. "Differential diagnostic patterns of dysarthria," J. Speech Hear. Res. **12**, 246–269.

Ebel, R. **(1951)**. "Estimation of the reliability of ratings," Psychometrica **16**, 407–424.

Fairbanks, G. **(1960)**. *Voice and Articulation Drillbook* (Harper & Row, New York).

Fritzell, B., Hammarberg, B., Gauffin, J., Karlsson, I., and Sundberg, J. **(1986)**. "Breathiness and insufficient vocal fold closure," J. Phon. **14**, 549–553.

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.

Hammarberg, B., Fritzell, B., Gauffin, J., Sundberg, J., and Wedin, L. **(1980)**. "Perceptual and acoustic correlates of abnormal voice qualities," Acta Oto-laryngol. **90**, 441–451.

Hirano, M. **(1981)**. *Clinical Examination of Voice* (Springer, Vienna).

Isshiki, N., Okamura, H., Tanabe, M., and Morimoto, M. **(1969)**. "Differential diagnosis of hoarseness," Folia Phoniatr. **21**, 9–19.

Isshiki, N., and Takeuchi, Y. **(1970)**. "Factor analysis of hoarseness," Studia Phonol. **5**, 37–44.

Jensen, P. J. **(1965)**. "Adequacy of terminology for clinical judgment of voice quality deviation," The Eye, Ear, Nose and Throat Monthly **44** (December), 77–82.

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.

Klich, R. J. **(1982)**. "Relationships of vowel characteristics to listener ratings of breathiness," J. Speech Hear. Res. **25**, 574–580.

Kreiman, J., Gerratt, B., Kempster, G., Erman, A., and Berke, G. **(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.

Kruskal, J., and Wish, M. **(1978)**. *Multidimensional Scaling*. Sage University Series on Quantitative Applications in the Social Sciences, series no. 07-011 (Sage, Beverly Hills, CA).

Ladefoged, P. **(1981)**. The relative nature of voice quality. Paper presented at the 101st Meeting of the Acoustical Society of America, Ottawa, Ontario.

Ladefoged, P., Maddieson, I., and Jackson, M. **(1988)**. "Investigating phonation types in different languages," in *Vocal Fold Physiology: Voice Production, Mechanisms and Functions*, edited by O. Fujimura (Raven, New York), pp. 297–317.

Laver, J. **(1980)**. *The Phonetic Description of Voice Quality* (Cambridge U. P., Cambridge).

Laver, J. **(1981)**. "The analysis of vocal quality: From the classical period to the 20th century," in *Toward a History of Phonetics*, edited by R. Asher and E. Henderson (Edinburgh U. P., Edinburgh), pp. 79–99.

Lively, M., and Emanuel, F. **(1970)**. "Spectral noise levels and roughness severity ratings for normal and simulated rough vowels produced by adult females," J. Speech Hear. Res. **13**, 503–517.

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.

Reed, C. G. **(1980)**. "Voice therapy: A need for research," J. Speech Hear. Disord. **45**, 157–189.

Sansone, F. Jr., and Emanuel, F. **(1970)**. "Spectral noise levels and rough-

ness severity ratings for normal and simulated rough vowels produced by adult males," J. Speech Hear. Res. **13**, 489–502.

Sapir, S., and Aronson, A. E. (**1985**). "Clinician reliability in rating voice improvement after laryngeal nerve section for spastic dysphonia," Laryngoscope **95**, 200–202

SAS Institute, Inc. (**1992**). "The MDS procedure," in SAS Tech. Rep. P-229, SAS/STAT Software: Changes and Enhancements, Release 6.07, SAS Institute, Inc., Cary, NC, pp. 251–286.

Schiffman, S., Reynolds, M., and Young, F. (**1981**). *Introduction to Multidimensional Scaling: Theory, Method, and Applications* (Academic, New York).

Shipp, T., and Huntington, D. (**1965**). "Some acoustic and perceptual factors in acute-laryngitic hoarseness," J. Speech Hear. Disord. **30**, 350–359.

Shrout, P., and Fleiss, J. (**1979**). "Intraclass correlations: Uses in assessing rater reliability," Psychol. Bull. **86**, 420–428.

Takahashi, H., and Koike, Y. (**1975**). "Some perceptual dimensions and acoustic correlates of pathological voices," Acta Oto-laryngol. Suppl. **338**, 2–24.

Titze, I., Horii, Y., and Scherer, R. C. (**1987**). "Some technical considerations in voice perturbation measurements," J. Speech Hear. Res. **30**, 252–260.

Wendahl, R. (**1966**). "Some parameters of auditory roughness," Folia Phoniatr. **18**, 26–32.

Whitehead, R. L., and Emanuel, F. W. (**1974**). "Some spectrographic and perceptual features of vocal fry, abnormally rough, and modal register vowel phonations," J. Commun. Disord. **1**, 305–319.

Wolfe, V., and Steinfatt, T. M. (**1987**). "Prediction of vocal severity within and across voice types." J. Speech Hear. Res. **30**, 230–240.

Yanagihara, N. (**1967**). "Hoarseness: investigation of the physiological mechanisms," Ann. Otol. Rhinol., Laryngol. **76**, 472–488.

Yost, W. A. (Ed.) (**1989**). *Classification of Complex Nonspeech Sounds* (National Academy, Washington, DC).

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.

Yumoto, E., Sasaski, Y., and Okamura, H. (**1984**). "Harmonics-to-noise ratio and psychophysical measurement of the degree of hoarseness," J. Speech Hear. Res. **27**, 2–6.