Journal of Speech and Hearing Research, Volume 35, 512–520, June 1992

Individual Differences in Voice Quality Perception

Jody Kreiman
Bruce R. Gerratt
Kristin Precoda
Gerald S. Berke
VA Medical Center, West Los Angeles
and
Division of Head and Neck Surgery
UCLA School of Medicine

Sixteen listeners (10 expert, 6 naive) judged the dissimilarity of pairs of voices drawn from pathological and normal populations. Separate nonmetric multidimensional scaling solutions were calculated for each listener and voice set. The correlations between individual listeners' dissimilarity ratings were low However, scaling solutions indicated that each subject judged the voices in a reliable, meaningful way. Listeners differed more from one another in their judgments of the pathological voices (which varied widely on a number of acoustic parameters) than they did for the normal voices (which formed a much more homogeneous set acoustically). The acoustic features listeners used to judge dissimilarity were predictable from the characteristics of the stimulus sets' only parameters that showed substantial variability were perceptually salient across listeners. These results are consistent with prototype models of voice perception They suggest that traditional means of assessing listener reliability in voice perception tasks may not be appropriate, and highlight the importance of using explicit comparisons between stimuli when studying voice quality perception

KEY WORDS: voice quality, perception (voice), agreement

The issue of precisely how and why listeners differ in their judgments of vocal quality has received little attention from researchers, despite the potential theoretical and clinical importance such differences may have. Many investigators apparently assume that all listeners share a single underlying set of perceptual features, and thus that different listeners' ratings of vocal quality should be in close agreement. In fact, levels of interrater reliability and agreement vary substantially across scales, listener groups, and voice sets (see Bassich & Ludlow, 1986, for review). For example, Deal and Emanuel (1978) reported 80% of ratings within ±1 scale value when 11 graduate students rated normal and simulated rough vowels on a 5-point equal-appearing interval scale. Eighty-eight percent of ratings were within ±1 scale value when pathological vowels were judged. Yumoto, Sasaki, and Okamura (1984) found correlations ranging from .51 to .79 when eight laryngologists rated the hoarseness of 87 voices on a 4-point equal-appearing interval scale. Bassich and Ludlow (1986) had four inexperienced but intensively trained listeners rate 10 pathological voices on 13 seven-point scales. They reported a mean intraclass correlation of .71, with a range of .19 to .96 across the 13 scales. Only three scales had intraclass correlations greater than .9. Based on these results and on the varying levels of interrater agreement reported in the literature, Bassich and Ludlow argued that listeners require substantial clinical experience in order to rate voices reliably (i.e., for their ratings to agree).

However, if listeners differ in perceptual strategies but they each judge voices in ways that are internally consistent and reasonable, then traditional standards for interrater reliability must be rethought. That listeners do in fact differ systematically from one another in their vocal quality judgments has been known for some time. Systematic differences between subjects were noted by Voiers (1964) in a factor analytic study of voice quality perception. He found both significant listener biases ("constant errors") and listener idiosyncracies (interactions between listeners and voice

0022-4685/92/3503-0512$01.00/0

samples). More recently, Kreiman, Gerratt, and Precoda (1990) used multidimensional scaling to examine the effects of listener groups and speaker populations on average perceptual strategies. Significant listener group and speaker effects were reported. Clinicians and naive listeners attended to different cues when judging the same voices, and listeners within groups differed in the vocal characteristics used to judge different sets of voices. Further, individual clinicians in this study differed significantly in the relative attention they paid to different acoustic parameters, whereas naive listeners did not. Kempster, Kistler, and Hillenbrand (1991) examined relative subject weights for two three-dimensional scaling solutions for 25 clinical trainees, who judged matched sets of dysphonic voices. They found good agreement about the relative importance of one dimension, with 21/25 listeners giving it the most weight for one set of judgments and 19/25 for the other set. However, their listeners did not agree about the relative importance of the remaining two dimensions. Further, roughly half their listeners weighed the three dimensions differently for the two sets of voices. Finally, R^2 values for a few listeners were quite low ($< .3$), suggesting these subjects differed substantially from the "average" perceptual strategy reflected in the group scaling analyses.

These previous studies demonstrate that individual listeners do deviate from an average perceptual strategy, at least with respect to the relative importance granted to different vocal features. However, it is unclear whether listeners in these studies attended to the same things when judging the quality of a given voice or if a given listener consistently relied on the same characteristics when judging the quality of different voices.

In the present study we undertook detailed analyses of the perceptual strategies of individual listeners. We wanted to determine what (if any) aspects of voice quality were perceptually important across listeners and voice sets, and what characteristics were unique to individual listeners and voice sets. If listener differences are limited to variations in the salience of a relatively constant set of perceptual dimensions, then voice perception models probably need not account for listener biases. Alternatively, if listeners do differ in which acoustic characteristics they exploit when judging vocal quality, then such differences must be taken into account when attempting to predict perceptual quality from the acoustic characteristics of a voice.

A second goal of this study was to determine if differences in listeners' perceptual judgments were systematically related to selected acoustic characteristics of the voices. If a fairly constant set of perceptual parameters is used for different sets of voices, then the perceptual context in which judgments are made may safely be ignored when predicting a voice's perceptual quality from its acoustic characteristics. Alternatively, if the acoustic correlates of perceived quality vary systematically with the context provided by other voices, this extra-stimulus factor also must be described as part of the voice perception process.

Method

Voice Selection and Recording Procedures

The stimuli used in this study have been described in detail elsewhere (Kreiman et al., 1990). Briefly, the voices of 18 male speakers with a variety of voice disorders and 18 normal male speakers were selected at random from a library of recordings. Speakers were recorded using a Brüel and Kjær condenser microphone and a high-fidelity, reel-to-reel tape recorder. They were asked to sustain /ɑ/ as long as possible with customary pitch and loudness.

Stimulus Tapes

Voice samples were low-pass filtered at 6.3 kHz and then digitized at 17.8 kHz using a 16-bit A/D converter installed in an IBM/AT-compatible computer. A 1.67-sec sample was taken from the most stable portion of each /ɑ/. Sample duration was determined by pilot tests and by hardware limitations. The digitized segments were normalized for peak voltage, and onsets and offsets were multiplied by a 10-msec ramp to help eliminate click artifacts.

Voices were output through a 16-bit D/A converter. Separate test tapes were constructed for the pathological and normal voice sets. Each tape included both orders (AB and BA) of all possible pairs of the 18 voices, for a total of 306 trials/voice set. Voice samples within a pair were separated by 1 sec. Pairs were separated by 6 sec.

The entire set of voice pairs was randomly ordered, with the constraint that the AB and BA orders of a pair of voices did not occur adjacent to one another. Stimuli were recorded on audio tape for presentation to listeners. All listeners heard the pairs in the same random order.

Listeners

Two listener groups participated in the experiment. The first (expert listeners) included eight speech pathologists and two otolaryngologists, each with a minimum of 2 years' experience evaluating and treating voice disorders. The second group (naive listeners) included six listeners with no training in speech pathology, linguistics, or audiology, and with no previous formal exposure to pathological voices.

Task

Each listener participated in two test sessions, one for each voice set. Sessions were held at least 1 week apart. At each session, listeners were first told that we were interested in how they as individuals judged each voice pair. They were asked to listen carefully to each pair, and to rate the dissimilarity[1] of the voices on a 7-point equal-appearing interval scale. Stimuli were presented in a sound-treated room at a constant listening level (approximately 80 dB SPL). To mimic normal clinical listening conditions, tapes were played in free field over two loudspeakers equidistant from the listener. Order of presentation of the voice sets was randomized across listeners. Each test session lasted approximately 1.5 hours.

[1]In dissimilarity measures, a large number indicates stimuli are very different. For similarity measures, a large number indicates great similarity.

Reliability of the Data

Because one purpose of this work was to assess the extent to which listeners agreed in their perception of the similarity of voice pairs, interrater agreement will be discussed with other results below. Issues surrounding intrarater agreement are discussed in the following section

Multidimensional Scaling Analyses

As described above, each listener produced two full 18 × 18 matrices (minus the diagonal) of dissimilarity judgments, one for each voice set. Many vowel perception studies (e.g., Macmillan, Goldberg, & Braida, 1988; Repp & Crowder, 1990), and data from studies of long- and short-term memory for voices (Kreiman & Gerratt, 1990; Kreiman & Papcun, 1991) show that in paired-comparison tasks the voices in a pair are not "equal." Rather, the first voice seemingly provides a context against which the second is judged. For example, voices that are both breathy and rough sound more breathy, and less rough, when presented in the context of a voice that is rough, but not breathy. Because we wanted our results to reflect these effects, both the AB and BA orders of the voice pairs were presented in this study. However, this decision also means that traditional measures of test-retest reliability (for example, the correlation between the first and second rating of a pair, or the number of ratings within ±1 scale value of one another) are not particularly appropriate for the present data set, because we *expect* ratings to vary with presentation order. The correlation between first and second ratings and the percentage of ratings within ±1 scale value are given in Table 1. As predicted, values are rather low overall, especially for the expert listeners.

Because presentation order is expected to affect dissimilarity ratings, the top and bottom halves of each matrix of data for each listener and voice set were analyzed together using the individual differences model of SAS PROC ALSCAL (SAS Institute, 1983; see also Schiffman, Reynolds, & Young, 1981). Output from the scaling program includes a measure of stress, which reflects the fit between the data and the scaling model, and R^2 values, which measure the amount of variance in the underlying data that is accounted for by the scaling solution. If differences between the first and second ratings of a pair of voices in fact represent noise in the data, low values of R^2 and high values of stress for the scaling solution will reflect this. Conversely, if such differences represent systematic trends in perceptual judgments, scaling solutions should account for most of the variance in the underlying data, and stress and R^2 values should reflect this fact.

Separate multidimensional scaling (MDS) analyses for each listener and voice set were conducted, for a total of 32 individual scaling solutions. Four additional group scaling analyses were also undertaken. Each included a single matrix of dissimilarity data (averaged across the diagonal) for every listener in a group, for a single voice set. These "group solutions" were used to determine the general perceptual trend across listeners and to determine the extent to which composite analyses reflect the strategies used by individual listeners.

Separate solutions in two, three, and four dimensions were found for each data set. Scaling solutions were selected based on interpretability and on R^2 and stress values. R^2 values for each solution are included in Tables 3 and 4 below.

Acoustic Measurements

Acoustic measures were obtained from each voice sample[2] for use in interpreting the derived perceptual dimensions. The fundamental frequency (F0) and the frequencies of the first three formants (F1, F2, and F3) were measured from displays on a digital spectrograph. For jitter and shimmer measurements, a waveform landmark (positive or negative peak or zero crossing) that could be identified reliably from cycle to cycle was selected by hand. Measurements of mean jitter, standard deviation of jitter, percent jitter, directional jitter, and the coefficient of variation for jitter were then calculated using parabolic interpolation when the point marked was a peak and using linear interpolation when a zero crossing was marked (Titze, Horii, & Scherer, 1987). Analogous shimmer measures were also calculated, using the difference in dB between the highest and lowest points between marked points as the amplitude. We also calculated the natural logarithm of the standard deviation of the period lengths (LNSD; see Wolfe & Steinfatt, 1987); the harmonics-to-noise ratio (HTN; see Yumoto, Gould, & Baer, 1982); the difference in the amplitudes of the first and second harmonics in dB (H1-H2), which has been associated with phonemic breathiness in languages that distinguish "clear" and "breathy" vowels (e.g., Bickley, 1982; Ladefoged, 1981); and

[2]The voice of one pathological speaker was clearly diplophonic, and only formant measurements were made for him, because assumptions of normal glottal (near-) periodicity required by other measures were not met (Gerratt, Precoda, Hanson, & Berke, 1988)

TABLE 1. Traditional measures of test-retest reliability for the naive and expert groups.

Group	Voice set	Mean Pearson's *r*	Range	Mean % ratings ±1 scale value	Range
Expert	Pathological	.48	23–.67	69.7	56.9–85.0
	Normal	.44	31–.58	69.6	50.3–83.7
Naive	Pathological	.56	47–.71	78.1	66.0–86.3
	Normal	.59	43–.77	85.2	73.9–92.2

the mean and standard deviation of the "partial period comparison" (PPC and PPC SD). The PPC was proposed by Ladefoged, Maddieson, and Jackson (1988; see also Kreiman et al., 1990) as a measure of phonemic breathiness, and is currently under evaluation in our laboratory as a measure of vocal roughness.

Results

Interrater Agreement and Reliability of the Data

Table 2 lists the mean correlations between unscaled similarity ratings for all possible pairs of raters within a group, for each set of voices. N for each correlation was 306, reflecting the 18 × 17 judgments made by each listener for each voice set. All pairwise correlations between listeners were significant at $p < .01$ (adjusted for multiple comparisons), but as the table shows, values were not particularly high, especially for the expert listeners. The mean correlation across voice sets and listener groups was only .45.

Despite apparently poor interrater agreement, in every case multidimensional scaling solutions for individual listeners accounted for most of the variance in the listener's dissimilarity judgments. For experts, R^2 values ranged from .67 to .93, and for naive listeners they ranged from .76 to .91 (Tables 3 and 4). Note that values for grouped data were generally lower than those for individual listeners, especially for the experts.

Interrater Variability in Perceptual Strategy

Multidimensional scaling solutions were interpreted by examining the correlation of the measured characteristics of the voices with stimulus coordinates on the dimensions produced by the scaling analysis. When several parameters were significantly correlated with a dimension, multiple regression was used to identify unique correlations between dimensions and parameters. Results are included in Tables 3 and 4, for pathological and normal voice sets, respectively.

For the pathological voices (Table 3), only three of the ten expert scaling solutions (for clinicians E1, E5, and E6) approximated the expert group solution. In two of these three (E5 and E6), the perceptual dimensions were weighed differently than in the group solution. F0 was less important, and H1-H2 was more important, to these listeners than to the average solution.

Perceptual spaces for the remaining expert listeners differed from the group space in various ways. Several spaces (for listeners E2, E8, E9, and E10) lacked dimensions for the PPC, H1-H2, or both. The spaces for listeners E3, E4, and E7 lacked dimensions from the group space and included dimensions not found in the group space.

Perceptual spaces for all six naive listeners differed substantially from the group perceptual space for the pathological voices. One space (for listener N3) lacked a dimension correlated with vowel formant frequencies, three spaces (N4, N5, and N6) were missing dimensions for both H1-H2 and formant frequencies, and the other spaces (N1 and N2) lacked a dimension for H1-H2 but included another dimension not found in the group scaling solution.

Thus for the pathological voices, some subgroups of subjects with similar perceptual strategies were found. However, in general, listeners differed substantially both from each other and from a group or "average" perceptual strategy. This pattern is consistent with the lower R^2 values for the group versus individual scaling solutions: combining listeners with rather different strategies in a group analysis generally results in lower R^2 values (e.g., see Wish, Deutsch, & Biener, 1972).

Listeners differed far less in the perceptual strategies used for the normal voice set. The perceptual spaces for 5 of the 10 experts (E5, E6, E8, E9, and E10) included the same dimensions as did the group space, although listener E6 weighed the dimensions differently than the group did. Spaces for four of the six naive listeners (N1, N3, N4, and N6) were fundamentally the same as the group space, although for all four, shimmer was relatively more important, and F0 relatively less important, than in the group space. All but one listener who differed from the group pattern lacked a shimmer dimension, and thus judged the similarity of the voices almost entirely in terms of F0. That one listener did attend to shimmer, but not to F0.

TABLE 2. Interrater reliability: Correlations between similarity ratings for pairs of expert and naive listeners.

Listeners	Voice set	Mean r (sd)	Range
Expert	Pathological	.48 (.09)	.27–.73
	Normal	.34 (.09)	.21–.55
Naive	Pathological	.53 (.06)	.42–.66
	Normal	.52 (.09)	.30–.70

Note. N = 306; all correlations significant at $p < .01$ (adjusted for multiple comparisons).

Intrarater Variability in Perceptual Strategy

A given listener did not necessarily use the same perceptual strategy when judging similarity for different voice sets. For example, consider the scaling solutions for listeners E5 and N5. For listener N5, F0 alone accounted for virtually all the explained variance in similarity ratings (81% out of a total of 90%) for the normal voices. F0 was also important for similarity judgments for the pathological voices for this listener, accounting for 58% of the variance in ratings; but vocal shimmer (which did not appear in this listener's solution for the normal voices) added an additional 26% to the explained variance.

For listener E5, F0 was also much more important for the normal than for the pathological voices (44% variance accounted for, vs. 21% for the pathological voices). Shimmer was nearly as important as F0 for the normal voices (33% variance accounted for), but apparently played almost no role in the perceptual space for the pathological voices. The PPC and H1-H2 were both important aspects of similarity for the pathological voices, but not for the normal voices, for this listener.

In fact, across scaling solutions, some acoustic parameters were consistently associated with dissimilarity judg-

TABLE 3. Interpretations of the individual multidimensional scaling solutions: Pathological voices.

Listener	Dim #	% Variance acct'd for*	Correlated with:	R^2 for regression
E1	1	33.2	PPC & F0 & H1-H2	.86
	2	28.2	F0 & Directional Jitter	.62
	3	17.0	PPC & H1-H2	69
E2	1	56.3	F0 & % Jitter	.94
	2	17.9	H1-H2 & Shimmer SD	54
E3	1	43.5	F0	.49
	2	29.5	LNSD & F1	.45
E4	1	43.3	H1-H2	.24
	2	29.5	Mean Shimmer & H1-H2	.76
E5	1	37.0	H1-H2 & Shimmer SD	.76
	2	26.3	PPC & H1-H2 & F0	79
	3	21.2	F0 & % Jitter	.66
E6	1	32.4	PPC & H1-H2	.64
	2	27.3	H1-H2 & HTN	.58
	3	24.3	F0	80
E7	1	33.6	F0 & Mean Shimmer	.57
	2	33.0	Shimmer Coeff. of Var	.32
E8	1	53.6	F0	.71
	2	22.7	F0	.30
E9	1	48.4	PPC & F0	.72
	2	36.5	F0 & % Jitter	.82
E10	1	50.3	F0 & Mean Shimmer	.75
	2	42.4	PPC	.36
Expert Group	1	29.5	F0	.88
	2	23.7	PPC	.68
	3	19.7	H1-H2	.51
N1	1	32.4	F0 & F2	.80
	2	26.1	HTN & F0	62
	3	25.7	LNSD & PPC & F3	86
N2	1	36.3	% Jitter & F0	.64
	2	24.6	F0	.50
	3	15.6	PPC	81
	4	12.8	F2 & Shimmer Coeff. of Var.	.65
N3	1	52.4	F0 & Directional Jitter	.79
	2	32.0	H1-H2 & F0	.75
N4	1	43.9	F0	.89
	2	23.2	Shimmer Coeff. of Var. & Jitter SD	.42
	3	15.9	F1 & Shimmer SD	.66
N5	1	58.3	F0	.90
	2	25.5	Shimmer Coeff. of Var.	.40
N6	1	66.0	F0	.82
	2	24.5	LNSD & PPC	.61
Naive Group	1	33.9	F0	.76
	2	25.3	F0 & Shimmer Coeff. of Var.	.72
	3	13.8	F1 & Directional Jitter	.49
	4	11.4	H1-H2 & Shimmer SD	.75

Note. PPC = partial period comparison; LNSD = natural logarithm of the standard deviation of the period lengths; HTN = harmonics-to-noise ratio; H1-H2 = the difference in the amplitudes of the first and second harmonics; Coeff. of Var. = the coefficient of variation (see text for details).
*For each listener, the sum of the percentage of variance accounted for by the individual dimensions equals R^2 for the entire scaling solution.

ments for the normal voices but not for the pathological voices, and vice versa. Table 5 shows the frequency with which individual acoustic characteristics of the voices emerged as significant correlates of perceptual dimensions across listeners. Recall that a total of 39 dimensions were extracted by the 16 scaling solutions for the pathological voice set, and 32 for the 16 normal voice solutions. In this table, a large number indicates that the parameter was significantly correlated ($r > .6$, $p < 0.01$, adjusted for multiple comparisons) with dimensions in perceptual spaces for many listeners. As this table shows, the PPC was significantly correlated with nine perceptual dimensions for the pathological voices, but with only one for the normal voices. In contrast, directional jitter was related to vocal dissimilarity for the normal voices, but not the pathological voices.

The observed variability in perceptual strategies across

TABLE 4. Interpretations of the individual multidimensional scaling solutions: Normal voices.

Listener	Dim #	% Variance acct'd for*	Correlated with:	R^2
E1	1	37.4	F0 & F2/F1	.82
	2	33.6	F3 & HTN	.36
E2	1	48.2	F0 & F1	.81
	2	24.9	F0 & F3	.48
E3	1	38.5	F0	.77
	2	28.2	Uninterpreted	
E4	1	40.4	Directional Jitter	.50
	2	29.6	Shimmer SD & HTN	.44
E5	1	44.3	F0	.79
	2	32.9	Shimmer Coeff. of Var.	.29
E6	1	41.3	Mean Shimmer	.67
	2	36.1	F0	.38
E7	1	47.6	F0 & F2/F1	.66
	2	23.4	Uninterpreted	
E8	1	49.8	F0	.87
	2	30.3	Mean Shimmer	.66
E9	1	53.0	F0 & Mean Shimmer	.91
	2	23.7	Mean Shimmer	.32
E10	1	36.3	F0 & Mean Shimmer	.75
	2	33.6	Shimmer SD	.53
Expert Group	1	30.3	F0	.88
	2	17.2	Shimmer SD & F3	.64
	3	15.5	Uninterpreted	
N1	1	38.4	F2/F1 & Mean Shimmer	.59
	2	37.7	F0	.88
N2	1	46.8	F0	.73
	2	31.0	F3 & F0	.61
N3	1	42.8	Shimmer SD & F2/F1	.62
	2	40.4	F0	.86
N4	1	53.1	Mean Shimmer	.62
	2	34.2	F0	.58
N5	1	81.2	F0	.98
	2	8.5	F3 & F1	.47
N6	1	46.0	Mean Shimmer & LNSD	.74
	2	43.2	F0	.90
Naive Group	1	61.4	F0	.96
	2	17.7	Shimmer Coeff. of Var. & F2/F1	.53

Note. PPC = partial period comparison; LNSD = natural logarithm of the standard deviation of the period lengths; HTN = harmonics-to-noise ratio; H1-H2 = the difference in the amplitudes of the first and second harmonics; Coeff. of Var. = the coefficient of variation (see text for details).
*For each listener, the sum of the percentage of variance accounted for by the individual dimensions equals R^2 for the entire scaling solution.

voice sets may be explained in part by differences in the acoustic characteristics of the voices in the sets. Table 5 lists the coefficient of variation associated with each acoustic parameter. The likelihood that a parameter will be perceptually salient to some listener is directly related to how much the voices vary on that parameter. With the exception of F0, which was important to virtually every listener, for both voice sets, only (but not all) parameters with a coefficient of variation greater than about 0.4 were significantly correlated with perceptual dimensions across listeners. Parameters that varied less were not related to perceptual judgments. As seen above, listeners did differ as to which parameters in this reduced set they attended to. However, parameters outside this set were not perceptually salient to any listener.

Finally, even listeners who shared common perceptual dimensions did not necessarily use that information similarly. Vocal fundamental frequency appeared in most scaling solutions; however, as Figure 1 shows, listeners differed in how they used F0 information. For example, F0 was used as a continuous dimension by some listeners, to sort voices into high- and low-pitched groups by others, and to sort voices into abnormal versus normal pitch groups by still others.

Summary and Discussion

Our findings may be summarized as follows. The expert group perceived the similarity of 18 pathological voices in terms of F0, the PPC, and H1-H2. F0 was the most important dimension in this group solution, accounting for nearly 30% of the variance in dissimilarity ratings. Individual clinicians varied considerably from this pattern, and no factor was

TABLE 5. The relationship between stimulus variability and perceptual salience for the pathological and normal voice sets.

Acoustic parameter	Coefficient of variation	# Perceptual dimensions significantly correlated with parameter ($r > .6$)
Pathological voices		
F3	0.11	0
F2	0.11	1
F2/F1	0.12	0
Directional Shimmer	0.14	0
F1	0 17	1
F2-F1	0.22	0
HTN	0.23	2
Directional Jitter	0 25	0
F0	0.29	21
Shimmer Coeff. of Var	0.40	3
Jitter Coeff. of Var	0.44	1
Mean Shimmer	0 45	7
PPC	0 48	9
LNSD	0.55	5
Shimmer SD	0.60	7
H1-H2	0.69	8
Mean Jitter	0 75	1
PPC SD	0.76	10
% Jitter	0.84	1
Jitter SD	0.92	1
Normal voices		
F3	0.07	0
F2	0.09	0
Directional Shimmer	0.09	0
F1	0.09	0
F2/F1	0 09	1
HTN	0.15	0
F0	0 16	21
F2-F1	0.19	0
Jitter Coeff. of Var.	0.20	1
H1-H2	0.26	0
PPC	0.34	1
PPC SD	0.41	11
% Jitter	0.48	9
Mean Shimmer	0.55	14
Shimmer Coeff. of Var.	0.59	1
Shimmer SD	0.61	14
Jitter SD	0.62	16
Mean Jitter	0.63	18
Directional Jitter	0.87	21
LNSD	1.95	14

Note. PPC = partial period comparison; LNSD = natural logarithm of the standard deviation of the period lengths; HTN = harmonics-to-noise ratio, H1-H2 = the difference in the amplitudes of the first and second harmonics; Coeff of Var. = the coefficient of variation (see text for details).

common to all individual solutions. Although F0 appeared in 9 of 10 perceptual spaces, listeners differed both in their use of F0 and in its relative perceptual importance.

The naive group perceived the pathological voices in terms of F0, F1, H1-H2, and perturbation. Although some individuals did differ from this pattern, differences were not as marked as for the expert group.

As a group, expert listeners perceived the normal voices in terms of F0, shimmer, and formant frequencies. Naive listeners relied mainly on F0. Although only a few listeners deviated from these general models, the fit of the scaling solutions to the underlying dissimilarity judgments was not as good for the normal as for the pathological voices. This is partly attributable to the fact that the scaling solutions for the normal voices all had two dimensions, while the solutions for the pathological voices ranged from two to four dimensions. Other things being equal, R^2 values increase with the number of dimensions extracted. This finding may also reflect differences in the way voice quality is judged for homogeneous (i.e., well matched) vs. heterogeneous sets of voices. When voices differ substantially in quality, as the pathological voices did, listeners have choices about what to listen to, but the features they select do an efficient job of distinguishing among speakers. When the voices in a set are relatively similar, as the normal voices were, listeners' strategies apparently converge on a relatively small set of perceptual features. However, these features account for a smaller proportion of the variance in dissimilarity ratings. The residual variance, we speculate, may reflect use of pair-specific criteria that do not apply uniformly across the entire voice set, analogous to Voiers' (1964) constant errors of interaction between listeners and voices. For any given listener, these varying criteria did not account for enough variance to emerge as full dimensions in a scaling analysis.

For the pathological voice set, clinicians differed more than naive listeners did. Both listener groups varied less in perceptual strategy for the normal than for the pathological voices. Clinicians apparently develop idiosyncratic approaches to rating pathological voice qualities in the course of their clinical training and practice. Extensive experience with pathologic voices may provide a richer (or noisier) auditory "palette" for subtle judgments regarding pathologic voice quality than that of naive listeners. Differences among experts' scaling solutions reflect these differences in individual palettes and in attentive strategies.

Many questions remain as to how voice perception should be modeled. Generally, perceptual models describe processes by which listeners map incoming signals onto abstract mental representations. Prototype models suggest that listeners use "typical" stimuli to organize their internal representations. In the case of speech, listeners apparently use prototypes corresponding to the phonemic categories of their language (see, e.g., Grieser & Kuhl, 1989). Since in English no "phonemic" categories exist for voices or phonation types, it is not obvious what information about voices is represented centrally. Listeners seem to use information about the central trend for a population of speakers when remembering an unfamiliar voice (Kreiman & Papcun, 1991; Papcun, Kreiman, & Davis, 1989). Additionally, listeners can make reliable (although not necessarily valid) judgments of a speaker's personality and mood (e.g., Allport & Cantril, 1934, Aronovitch, 1976; see Kramer, 1963, or Scherer, 1979 for review), age (e.g., Shipp & Hollien, 1969), and sex (e.g., Coleman, 1976; Pear, 1931) from voice quality. It is possible that expert listeners also store "typical" exemplars for various perceptual qualities (e.g., rough voices or breathy voices). Both the number and nature of qualities so represented probably vary across listeners, although evidence from the perception of musical timbre suggests no more than a few qualities at most are used (Bloothooft & Plomp, 1988).

The particular processes by which listeners map incoming signals onto representations apparently vary with attention

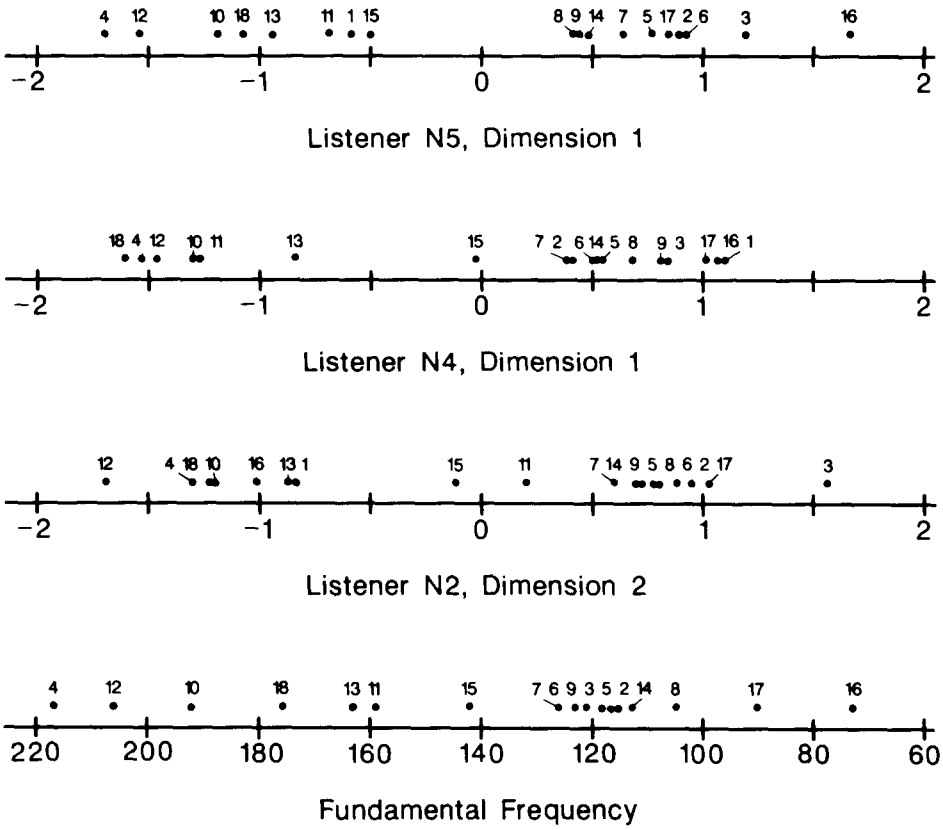

FIGURE 1. Multidimensional scaling dimensions corresponding to F0, for three listeners. Speakers (identified by number) are arranged by their fundamental frequency on the bottom trace, and by coordinates on perceptual dimensions in the upper three traces. The dimension for listener N5 (top trace) is nearly linearly related to F0; that is, the distance between voices on this dimension matches the distance between them in F0. Listener N4 (second trace) clustered voices into high- and low-pitched groups; differences between this listener and listener N5 are especially apparent for high-pitched voices. Listener N2 (third trace) grouped abnormally high- and low-pitched voices together, with voices in the normal pitch range clustered together. Speaker 1 was diplophonic, and F0 was thus undefined for him. Note the varying placement of this voice on the different dimensions.

and with context. Our findings suggest that three principles govern perceptual processing in the dissimilarity judgment task used here:

1. For any voice set, an acoustic parameter's potential perceptual salience can be estimated from its variability. If the coefficient of variation for a parameter is greater than about 0.4, the parameter may be exploited by listeners; if less, the parameter will not be perceptually important.

2. The more heterogeneous the voice set, the more heterogeneous the perceptual strategies that will be applied to it.

3. F0 is an exception to the above principles and is always important for voice quality judgments.

It is possible that F0 would fit the pattern shown by the other acoustic parameters if variability were measured in perceptually based units (e.g., JNDs) rather than with the coefficient of variation. Unfortunately, the relevant psychometric data are not available for most of the measures used here. Other parameters (intensity in particular; see Kempster et al., 1991, or Voiers, 1964) may also be important across voice sets in nonexperimental situations where stimuli are not matched or equalized for loudness.

The above principles highlight the importance of gathering perceptual judgments using explicit comparisons among voices whenever possible. Many commonly used perceptual tasks (e.g., direct magnitude scaling and long-term memory tasks) involve comparing voices to stored representations (i.e., internal standards or memory traces). These standards may fluctuate from trial to trial, and in any event are not directly available to the listener or the investigator. Consequently, calibration of a listener's behavior is not possible. Under some circumstances this is unavoidable. For example, in forensic situations a standardized set of comparison stimuli generally will not be available.

However, for clinical voice evaluations it is possible to develop protocols that use explicit comparisons between the patient's voice and a standard set of synthetic or natural comparison voices. The paired-comparison task used in this study is cumbersome for clinical use, as it requires listeners to make too many perceptual judgments. However, protocols using anchor stimuli, which provide a fixed, standardized context for voice quality assessment, could be developed to help control context-related inter- and intrarater variability in voice quality ratings. Based on the results presented here,

we hypothesize that controlling the acoustic context in which quality judgments are made will greatly increase the reliability of clinical voice evaluations. Further, knowledge of the acoustic (and other) characteristics of these explicit voice standards will permit us to determine how listeners' perceptual strategies vary across tasks, occasions, and voices.

In conclusion, listeners in this experiment differed markedly in their perception of the same voices and attended to different parameters when judging pathological and normal stimuli. Although these data do not meet traditional standards of "reliability," it would be wrong to conclude that the data are noisy or unreliable, because simple, meaningful interpretations were found to account for most of the observed variance. Instead, these results argue strongly for more careful investigation into the sources of variability in voice quality perception. A more detailed and accurate understanding of this variability will help unravel the many questions that remain about how listeners evaluate voices and will lead to standard protocols for clinical voice evaluations.

Acknowledgments

We thank Andrew Erman, Patty Gomeztrejo, David Hanson, Jean Holle, Daniel Kempler, Linda Mackey, Cynthia Moreno, and Jill Zweier for serving as expert listeners. We also thank Andrew Erman (again!) and Justin Woo for their help with data analysis. Gail Kempster, Alan Reich, and an anonymous reviewer provided many helpful comments on an earlier version of this paper. This research was supported by NIDCD award NS20707, by a NIDCD postdoctoral traineeship to the first author (NS07059), and by Veterans' Administration Rehabilitation Research and Development funds.

References

Allport, G. W., & Cantril, H. (1934). Judging personality from the voice. *Journal of Social Psychology, 5,* 37–55.

Aronovitch, C. D. (1976). The voice of personality: Stereotyped judgments and their relation to voice quality and sex of speaker. *Journal of Social Psychology, 99,* 207–220.

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

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

Bloothooft, G., & Plomp, R. (1988). The timbre of sung vowels. *Journal of the Acoustical Society of America, 84,* 847–860.

Coleman, R. O. (1976). A comparison of the contributions of two voice quality characteristics to the perception of maleness and femaleness in the voice. *Journal of Speech and Hearing Research, 19,* 168–180.

Deal, R. E., & Emanuel, F. W. (1978). Some waveform and spectral features of vowel roughness. *Journal of Speech and Hearing Research, 21,* 250–264.

Gerratt, B. R., Precoda, K., Hanson, D. G., & Berke, G. S. (1988, May). *Source characteristics of diplophonia.* Paper presented at the 115th Meeting of the Acoustical Society of America, Seattle, Washington.

Grieser, D., & Kuhl, P. (1989). Categorization of speech by infants. Support for speech sound prototypes. *Developmental Psychology, 25,* 577–588.

Kempster, G. B., Kistler, D. J., & Hillenbrand, J. (1991). Multidimensional scaling analysis of dysphonia in two speaker groups. *Journal of Speech and Hearing Research, 34,* 534–543

Kramer, E. (1963). Judgment of personal characteristics and emotions from nonverbal properties of speech. *Psychological Bulletin, 60,* 408–420.

Kreiman, J., & Gerratt, B. R. (1990, November). *Multidimensional perceptual spaces for vocal breathiness and roughness.* Paper presented at the 120th Meeting of the Acoustical Society of America, San Diego, California.

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

Kreiman, J., & Papcun, G. (1991). Comparing discrimination and recognition of unfamiliar voices. *Speech Communication, 10,* 265–275.

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

Ladefoged, P., Maddieson, I., & Jackson, M. (1988). Investigating phonation types in different languages. In O. Fujimura (Ed.), *Vocal fold physiology: Voice production, mechanisms and functions* (pp. 297–317). New York: Raven Press.

Macmillan, N., Goldberg, R., & Braida, L. (1988). Resolution for speech sounds: Basic sensitivity and context memory on vowel and consonant continua. *Journal of the Acoustical Society of America, 84,* 1262–1280.

Papcun, G., Kreiman, J., & Davis, A. (1989). Long-term memory for unfamiliar voices. *Journal of the Acoustical Society of America, 85,* 913–925.

Pear, T. H. (1931). *Voice and personality as applied to radio broadcasting.* New York: John Wiley and Sons.

Repp, B., & Crowder, R. (1990). Stimulus order effects in vowel discrimination. *Journal of the Acoustical Society of America, 88,* 2080–2090.

SAS Institute, Inc. (1983). *SUGI supplemental library user's guide.* Cary, NC: SAS Institute, Inc.

Scherer, K. R. (1979). Personality markers in speech. In K. Scherer and H. Giles (Eds.), *Social markers in speech* (pp. 147–210). Cambridge: Cambridge University Press.

Schiffman, S., Reynolds, M., & Young, F. (1981). *Introduction to multidimensional scaling: Theory, method, and applications.* New York: Academic.

Shipp, T., & Hollien, H. (1969). Perception of the aging male voice. *Journal of Speech and Hearing Research, 12,* 703–710.

Titze, I., Horii, Y., & Scherer, R. C. (1987). Some technical considerations in voice perturbation measurements. *Journal of Speech and Hearing Research, 30,* 252–260.

Voiers, W. D. (1964). Perceptual bases of speaker identity. *Journal of the Acoustical Society of America, 36,* 1065–1073.

Wish, M., Deutsch, M., & Biener, L. (1972). Differences in perceived similarity of nations. In N. Shepard, A. Romney and S. Nerlove (Eds.), *Multidimensional scaling: Theory and applications in the behavioral sciences,* vol. 2. New York: Seminar Press.

Wolfe, V., & Steinfatt, T. M. (1987). Prediction of vocal severity within and across voice types. *Journal of Speech and Hearing Research, 30,* 230–240.

Yumoto, E., Gould, W. J., & Baer, T. (1982). Harmonics-to-noise ratio as an index of the degree of hoarseness. *Journal of the Acoustical Society of America, 71,* 1544–1550.

Yumoto, E., Sasaki, Y., & Okamura, H. (1984). Harmonics-to-noise ratio and psychophysical measurement of the degree of hoarseness *Journal of Speech and Hearing Research, 27,* 2–6

Received January 7, 1991
Accepted July 18, 1991

Contact author: Jody Kreiman, PhD, West Los Angeles VAMC, Audiology and Speech Pathology (126), Wilshire and Sawtelle Boulevards, Los Angeles, CA 90073.