Journal of Speech and Hearing Research, Volume 36, 14–20, February 1993

Comparing Internal and External Standards in Voice Quality Judgments

Bruce R. Gerratt
Jody Kreiman
Norma Antonanzas-Barroso
Gerald S. Berke
VA Medical Center, West Los Angeles
and UCLA School of Medicine
Los Angeles, CA

A new descriptive framework for voice quality perception (Kreiman, Gerratt, Kempster, Erman, & Berke, 1993) states that when listeners rate a voice on some quality dimension (e.g., roughness), they compare the stimulus presented to an internal standard or scale. Hypothetically, substituting explicit, external standards for these unstable internal standards should improve listener reliability. Further, the framework suggests that internal standards for vocal qualities are inherently unstable, and may be influenced by factors other than the physical signal being judged. Among these factors, context effects may cause drift in listeners' voice ratings by influencing the internal standard against which judgments are made. To test these hypotheses, we asked 12 clinicians to judge the roughness of 22 synthetic stimuli using two scales: a traditional 5-point equal-appearing interval (EAI) scale and a scale with explicit anchor stimuli for each scale point. The stimulus set included a relatively large number of normal and mildly rough voices. We predicted that this would produce an increase in the perceived roughness of moderately rough stimuli over time for the EAI ratings, but not for the explicitly anchored ratings. Ratings made using the anchored scale were significantly more reliable than those gathered using the unanchored paradigm. Further, as predicted, ratings on the unanchored EAI scale drifted significantly within a listening session in the direction expected, but ratings on the anchored scale did not. These results are consistent with our framework and suggest that explicitly anchored paradigms for voice quality evaluation might improve both research and clinical practice.

KEY WORDS: voice quality, perception (voice), ratings (voice), reliability

Perception of a patient's voice is at the heart of evaluating and treating patients with voice disorders. Patients and their families decide whether treatment has been successful based largely on whether the patient sounds better. Similarly, clinicians make many decisions about managing speech and voice disorders based upon perceptual judgments. Indeed, a recent survey of speech/language pathologists in the VA medical system demonstrated their overwhelming preference for perceptually based over instrumental measures of speech and voice (Gerratt, Till, Rosenbek, Wertz, & Boysen, 1991). This central role of voice quality perception is not surprising, considering that the goal of speech is communication. It follows that the ultimate test of the acceptability of speech should involve its perceptual acceptability to listeners (Moll, 1964).

The importance of perceptual measures is also demonstrated by their frequent use as a standard against which acoustic measures are validated or compared. Researchers proposing objective voice measures often demonstrate their measure's utility by reporting a correlation between the measure and ratings of perceived vocal quality (e.g., Fukazawa & El-Assuooty, 1988; Hillenbrand, 1988; Klatt & Klatt, 1990; Kojima, Gould, Lambiase, & Isshiki, 1980; Ladefoged, Maddieson, & Jackson, 1988;

0022-4685/93/3601-0014$01.00/0

Takahashi & Koike, 1975). Thus, listener judgments are essential, both for clinical consideration and criterion validation of instrumental voice measures.

Despite their importance, these "subjective" measures of voice quality are not highly regarded as either clinical or research tools because of inherent problems with intra- and interjudge reliability (e.g., Cullinan, Prather, & Williams, 1963; Ludlow, 1981), because they are considered to lack objectivity and do not require great technical sophistication (Weismer & Liss, 1991), and because there is no accepted set of perceptual scales used by clinicians (e.g., Jensen, 1965; Yumoto, Gould, & Baer, 1982). In part because of these views, so-called "objective," nonperceptual measures for vocal assessment have received much more attention in voice research.

Thus, a paradox exists in the study of voice quality. Serious concerns regarding judgment reliability and uncertainty regarding the use and meaning of various rating scales have led some to abandon perceptual measures in favor of instrumental approaches to vocal assessment. However, as pointed out by Moll (1964), if a measure of vocal quality is to be useful, it must be closely related to listener judgments of that voice quality dimension. Both clinical and research practices are built upon perceptual data, but these data have never been gathered in ways that foster the confidence of clinicians or researchers.

This apparent contradiction has resulted in part from a lack of cogent research into the sources of variability in voice quality judgments. A review of the literature in a companion study (Kreiman, Gerratt, Kempster, Erman, & Berke, 1993) argues that most research utilizing perceptual ratings of voice quality assumes that voice ratings comprise only two components: the acoustic voice signal being rated and random error. Kreiman et al. propose an alternative descriptive framework specifying several other sources of variability in voice ratings. The new framework states that when listeners rate a voice on some quality dimension (e.g., roughness), they compare the stimulus presented to an internal standard or scale. These internal standards are developed out of a listener's experience with voices and are maintained in memory; accordingly, they differ from listener to listener. Further, internal standards for vocal qualities are inherently unstable and may be influenced by internal factors, such as lapses in memory and attention, and external variables, such as acoustic context (Kreiman, Gerratt, Precoda, & Berke, 1992) and listening task.

The descriptive framework proposes that an observed voice rating includes variability associated with differences among listeners in experience, in overall sensitivity to the characteristic being rated, and/or in response bias (see Kreiman, Gerratt, & Precoda, 1990; Kreiman et al., 1992). Additionally, effects related to the specific rating task contribute to rating variability. These include context effects and the number of points on the rating scale (Kreiman et al., 1993; Rossi, Pavlovic, & Espesser, 1990). Interactions between task and listener factors may also occur.

This framework implies that traditional voice-rating procedures are akin to matching tasks, because rating stimuli involves matching them to stored mental representations. Thus, the framework suggests that variability in voice quality ratings might be reduced by replacing idiosyncratic, unstable, internal standards with fixed external standards. A protocol using fixed reference voices would control for differences among listeners in their internal standards for different voice qualities by providing all raters with a constant set of referents. This protocol would also control context-related variability, because external standards remain constant from trial to trial. If our view is correct, controlling these two major sources of rating error should result in voice ratings that are significantly more reliable than those gathered using traditional paradigms.

As a preliminary test of this hypothesis, we asked expert listeners to rate the vocal roughness of synthetic voices using two protocols. The first was a traditional 5-point equal-appearing interval (EAI) scale. The second was a 5-point scale with each scale value represented by a synthetic voice sample, or anchor. In the EAI task, listeners judged the synthetic stimuli against their own internal standards for roughness. In the "anchored" task, they made their judgments against explicit, external standards for each scale value. We hypothesized that replacing idiosyncratic, unstable internal standards with fixed external standards would increase both intra- and interrater reliability for these ratings.

To test the prediction that internal standards can be influenced in consistent ways by the context in which judgments are made, we included a large number of normal and mildly abnormal voice samples in the stimulus set. We predicted that in the EAI task ratings of moderately rough voices would grow more severe as the test session progressed, because these voices would sound increasingly rough over time in a context weighted with mildly rough voices. We predicted that drift would not occur in the anchored task, where the explicit anchors were not subject to the influence of context.

Method

Synthesis Techniques

Twenty-two tokens of /ɑ/, each 1.5 sec long, were synthesized using the C language version of the Klatt MITalk speech synthesizer (Allen, Hunnicutt, & Klatt, 1987; Klatt & Klatt, 1990). The impulse voicing source was used. For all tokens, F0 was set at 125 Hz, F1 was 620 Hz, F2 was 1220 Hz, and F3 was 2550 Hz. The synthesizer parameters AV (amplitude of normal voicing source) and AH (amplitude of aspiration noise source) were systematically varied across the 22 stimuli. The first synthetic vowel used the default parameters for AV and AH (60 and 0, respectively) and represented "normal" voice quality. For subsequent stimuli, parameter AV was stepped from 60 to 49, and parameter AH was stepped from 38 to 61. Pilot studies showed that the difference between AH values of 0 and 38 was the minimum step that produced a reliably perceptible difference in voice quality. Default values were used for all other synthesizer parameters.

Stimuli were synthesized at a sample rate of 10 kHz. All stimuli began and ended with 50 msec ramps to eliminate sharp onsets and to increase naturalness. AV and AH values

TABLE 1. Synthesizer control parameters for the 5 anchor stimuli.

Stimulus number	AV	AH
1	60	0
9	57	45
13	55	49
17	53	53
21	51	57

were set at 30 at stimulus onset; F0 began at 90 Hz. Values for all three parameters then increased linearly to the appropriate target value; they declined linearly from the target value over the last 50 msec of the stimuli. Stimuli were equalized for peak amplitude prior to playback.

Anchor Stimuli

Five stimuli were chosen from the synthetic continuum to serve as examples, or "anchors," for each of the five scale points. Synthesizer parameters for these five voices are listed in Table 1. Magnitude spectra and waveforms for the first, third, and fifth anchor stimuli are given in Figure 1. Anchor stimuli were selected so that (a) they were discriminable with 100% accuracy in pilot tests; (b) they spanned the entire range of roughness represented by the synthetic stimuli; (c) they were approximately perceptually equidistant, as judged by the authors; and (d) at least three voices on the continuum separated each pair of anchors.

Listeners

Twelve listeners participated in this study. Three were speech pathologists and nine were otolaryngologists. All had a minimum of 2 years' experience evaluating vocal pathology ($M = 3.9$; $SD = 2.28$). All reported normal hearing.

Procedure

Ratings were gathered using two protocols. The first was a traditional 5-point equal-appearing interval (EAI) scale, where 1 represented normal voice quality and 5 represented severe roughness. The second was a 5-point scale with each scale value represented by a synthetic voice sample. Again, the sample labeled "1" was normal, and that labeled "5" was severely rough. Each listener participated in two listening sessions, one for each task. Sessions were separated by at least one week. Order of task presentation was randomized across subjects.

Stimuli were output through a 16-bit D/A converter, lowpass filtered with a 4-pole Bessel filter at 3 kHz, and presented in free field at approximately 80 dB SPL. Because most listeners were unfamiliar with synthetic speech, the entire set of stimulus voices was played twice in sequence (from most normal to most abnormal) at the beginning of each listening session. For the EAI task, listeners were instructed to concentrate on roughness, to ignore any other characteristics the voice might have, and to make their judgments with respect to their own criteria for roughness. Listeners were able to play the stimuli as often as needed before making their responses. For the anchored task, they were told to select the anchor that best matched the stimulus voice in its level of roughness. Again, they were instructed to ignore other characteristics of the voices. The subjects were able to play the stimuli and the anchors as often as needed before responding.

Stimuli were rerandomized for each listener. For both protocols, listeners rated the stimulus set twice in succes-

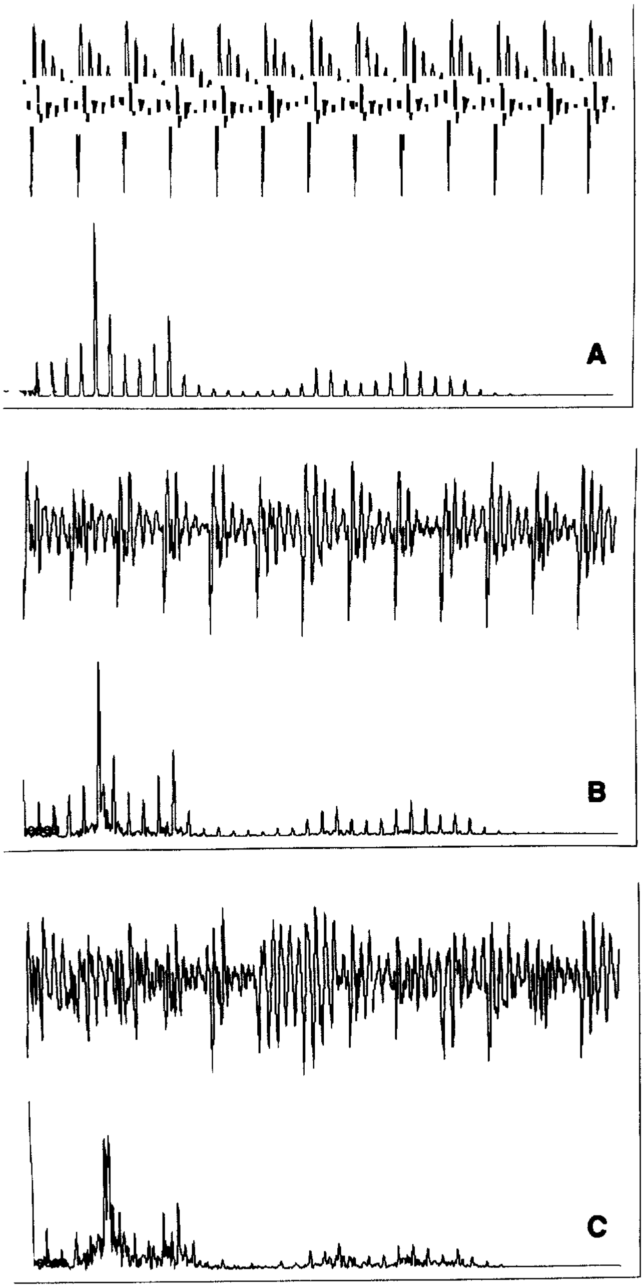

FIGURE 1. 100 msec waveforms (top) and respective magnitude spectra (bottom) for selected anchor stimuli. A: "normal" voice quality; B: moderate roughness (level 3); C: severe roughness (5 on the rating scale).

TABLE 2. Rating reliability for the EAI and anchored tasks.

	Intrarater reliability		Interrater reliability	
	EAI scale	Anchored scale	EAI scale	Anchored scale
Mean % exact agreement	54.2	75.0	50.2	68.4
SD	9.57	12.33	8.88	12.54
Range	40.9 – 68.2	50 – 90.9	27.3 – 70.5	36.4 – 88.6
Mean % ± 1 scale value	94.0	99.6	92.8	98.7
SD	7.07	1.30	5.22	2.43
Range	81.8 – 100	95.5 – 100	77.3 – 100	90.9 – 100
Mean Pearson's *r*	.86	.95	.85	.93
SD	.08	.03	.06	.03
Range	.71 – .95	.89 – .98	.71 – .94	.84 – .98
Intraclass correlation			.90	.95
95% Confidence intervals			.83 $<\rho<$ .95	.92 $<\rho<$ .97

sion, in two different random orders. Listeners were not informed that stimuli were repeated within a session.

Results

Rating Validity

Voice ratings for the two tasks were highly correlated for each listener. Values of Pearson's *r* ranged from .82 to .96, with a mean correlation of .89 and a standard deviation of .05. This suggests that the two rating paradigms captured essentially the same information about the stimuli.

Average Reliability and Agreement

The average intra- and interrater reliability of the 12 listeners is given in Table 2. Assuming all scale values are equally likely, the probability of responding within ±1 scale value by chance is .52 for both procedures. The chance probability for obtaining identical responses is .20. The intraclass correlation (ICC; Ebel, 1951) was calculated using a two-way ANOVA to assess the reliability of a single rating (model [2,1]; see Shrout & Fleiss, 1979). A mixed-model ANOVA was used, treating voices as a fixed effect and listeners as a random effect. The ICC measures the overall cohesiveness of a group of raters (vs. the comparisons of pairwise measures above) and reflects the extent to which the present data might generalize to a new random sample of listeners. Confidence intervals about the ICC were calculated using the formula in Shrout and Fleiss (1979).

Reliability and agreement levels for both tasks were high on the average. As predicted, both intrarater and interrater reliability and agreement were substantially better for the anchored paradigm than for the EAI scale. The two tasks differed significantly in intrarater reliability on all measures. Exact intrarater agreement averaged 21% higher for the anchored task (matched pairs t-test; $t = -5.24$, $df = 11$, $p < .05$ one-tailed), and agreement within ±1 scale value averaged 6% higher (matched pairs t-test; $t = -2.61$, $df = 11$, $p < .05$ one-tailed). Ceiling effects limited the extent of differences between tasks on this measure. Within-listener values of Pearson's *r* averaged .09 higher for the anchored task (matched pairs t-test; $t = -3.66$, $df = 11$, $p < .05$ one-tailed).

Interrater differences between tasks were also significant for all agreement/reliability measures comparing pairs of listeners (Pearson's *r*, % ratings that agreed exactly or within ± one scale value). Exact agreement among pairs of raters averaged 18% better for the anchored task (matched pairs t-test; $t = -11.48$, $df = 65$, $p < .05$ one-tailed); agreement within ± one scale value averaged 6% better (matched pairs t-test; $t = -11.36$, $df = 65$, $p < .05$ one-tailed). Values of Pearson's *r* averaged .08 higher for an anchored task (matched pairs t-test; $t = 11.00$, $df = 65$, $p < .05$ one-tailed). The 95% confidence intervals for the intraclass correlation indicate that the two tasks did not differ significantly in reliability on this measure, largely because the confidence interval for the EAI task is quite wide. Thus, comparisons of pairs of raters indicated that any two listeners will agree significantly better on the anchored than on the EAI paradigm, but large differences in rating patterns among listeners on the EAI task prevented an overall measure from reaching significance.

Confidence Intervals for Ratings of Individual Voices

Figure 2 shows the width of the 95% confidence interval for the mean rating of each voice, plotted against the mean rating for that voice. The larger the confidence interval, the more variable the rating of the voice.

Confidence intervals are wider overall for the EAI task than for the anchored task ($t = 7.63$, $df = 21$, $p < .05$ one-tailed), reflecting the greater overall variability of EAI ratings. Further, patterns of error are different for the two tasks. The curve for the EAI ratings resembles that observed for natural stimuli (e.g., Kearns & Simmons, 1988; Kreiman et al., 1993), with better agreement among listeners for normal and extremely rough stimuli, and less agreement about moderately rough voices. However, for the anchored task ratings were less variable for stimuli identical to anchors (filled circles in

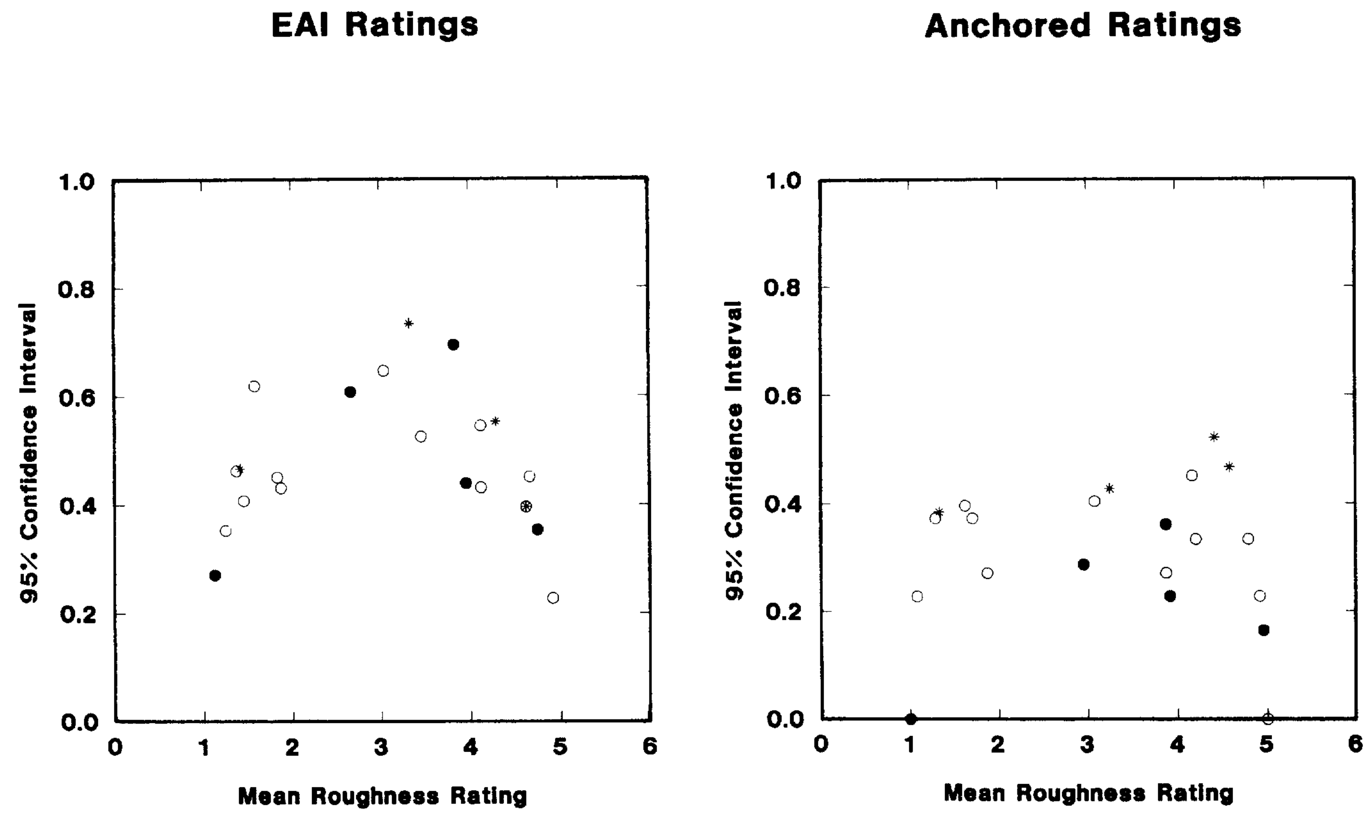

FIGURE 2. Confidence intervals for ratings of individual voices versus the mean rating for that voice. Larger confidence intervals indicate more variability among ratings. A: EAI task. B: Anchored task. Confidence intervals for ratings of stimuli that are identical to anchors are represented by filled circles, and confidence intervals for ratings of stimuli located exactly between anchors are represented by stars. Confidence intervals for ratings of the remaining stimuli are plotted with open circles.

the figure) and most variable for stimuli located exactly between anchors (stars in the figure).

Context Effects

Ratings of moderately rough stimuli drifted significantly from rating to re-rating for the EAI task, but not for the anchored task. Voices rated 2, 3, or 4 (i.e., not normal or severe) when initially presented sounded rougher the second time they were rated for the EAI task ($t = -2.10$, $df = 153$, $p < .05$ one-tailed). No significant difference in first versus second ratings was observed for moderately rough voices when they were rated on the anchored scale ($t = -0.16$, $df = 143$, $p > .05$ one-tailed).

Discussion

Levels of intra- and interrater agreement and reliability were very high overall in this study, both for the EAI and the anchored task. This is due in part to the relative simplicity of our synthetic stimuli, compared to the perceptual complexity of natural pathological voices. The synthetic voices varied along a single dimension and were free of other perceptual qualities that might have complicated the rating process. Thus on both tasks listeners were limited in the extent to which they could selectively focus on different aspects of the voices or vary their mental definitions of roughness. Both reliability levels and listener comments indicated that rating the roughness of these voices was easy, on either the anchored or EAI scale. The high correlation between the two sets of ratings further indicated that both tasks captured the same information about the stimuli, again possibly due to the limited scope these stimuli offered for differences in perceptual strategy to emerge.

Despite possible ceiling effects, both intra- and interrater reliability were significantly greater for the anchored paradigm than for the EAI paradigm, as predicted by our descriptive framework. In this framework, both tasks are viewed as involving matching stimuli to standards. In the case of the EAI task, these standards are internal, unobservable, and unstable. In the externally anchored task, they are explicit and constant.

As further predicted by our framework, ratings of moderately rough voices drifted significantly from the first to the second presentation within the test session in the EAI task, but not in the anchored task. Drift occurred in a context weighted with normal and mildly rough voices, presumably because that context modified a listener's internal standards for roughness.

Patterns of error differed for the two tasks in other ways as well. For the EAI task, agreement was better for normal and extremely rough stimuli, and worse for moderately rough stimuli. For the anchored task, listeners agreed better about the roughness of voices located near the anchors, and less well about voices located between anchors. Similar findings have been reported for judgments of frequency (Pollack, 1953) and intensity (Berliner, Durlach, & Braida, 1978) of tones, and for ratings of the naturalness of synthetic speech (Rossi, Pavlovic, & Espesser, 1990). In the studies of frequency and intensity, presentation of a single anchor drawn from the stimulus continuum improved listeners' ability to identify the stimuli in the vicinity of the anchor. Consistent with our findings, improvement was greatest when the anchor was drawn from the middle of the continuum, and was minimal when it was drawn from either extreme.

These patterns of variability suggest that "scale resolution" may contribute to variability in voice quality ratings on the anchored paradigm. If anchor stimuli are perceptually far apart, listeners may perceive some voices as being located exactly between anchors. In this case, voices located near anchors will be consistently rated, and voices equidistant between anchors will vary in their ratings. Thus, measurement error will be concentrated in particular regions of a scale, rendering ratings of some voices much more reliable than ratings of other voices that may not be very different acoustically (see Figure 2). Conversely, if anchors are too close together they may not be reliably distinguishable, and listeners may treat them as interchangeable. In this situation, overall levels of error increase, but error varies smoothly along the entire rating scale, leading to patterns like those found for the EAI task. Measuring voices using any rating scale will always involve a compromise between these two kinds of error. Thus, perfectly reliable ratings are not possible, even in theory. Figure 2 suggests that the 5-point scale is too coarse relative to listeners' average sensitivity to differences among stimuli. At least for the stimuli used here, a 7-point scale would probably represent a better compromise between errors due to widely and narrowly separated anchors. This suggestion remains to be investigated.

The present findings are consistent with the proposal that listeners' internal standards for a vocal quality are relatively stable for extreme qualities (normal or severe), but that judgments of intermediate levels may be influenced by factors other than the magnitude of the quality being judged. We speculate that as a listener hears many (near-) normal voices, the listener's internal standard(s) for a quality moves toward the normal end of the quality dimension. The distance between the standard(s) and a moderately rough voice thus increases, and the voice appears rougher as a result. This process has analogues in other sensory modalities. For example, a constant room temperature feels warmer after prolonged exposure to the cold, or feels cooler after exposure to warmer temperatures. Our framework predicts that a context weighted with severely rough stimuli would cause internal standards to drift in the other direction, making voices sound less rough. This hypothesis also remains to be tested.

In the context of our framework, the relative stability of EAI ratings of extreme qualities has two possible explanations. Ratings of these voices may be stable because listeners' internal standards for normal and extreme qualities are well-developed and stable. Braida and his colleagues (e.g., Braida & Durlach, 1972; Berliner, Braida & Durlach, 1977) have described similar findings in intensity perception as "edge effects" and attribute them to the stability of internal reference points for extreme stimuli (Braida, Lim, Berliner, Durlach, Rabinowitz, & Purks, 1984).

Alternatively, the relative stability of extreme ratings may reflect the fact that the anchored scale and the EAI scales typically used in voice research have fixed endpoints and thus that ratings for voices initially perceived as normal or severe have nowhere to go. In the present case, voices that were initially rated 5 may indeed have seemed rougher on second presentation, but no 6 was available to raters. A task that does not limit the maximum or minimum rating a voice may receive (for example, direct magnitude estimation; see e.g., Cullinan et al., 1963; Emanuel & Smith, 1974; Schiavetti, Metz, & Sitler, 1981) could be used to select between these explanations.

Several factors limit the conclusions that may be drawn here. First, the synthetic stimuli varied from one another along only one dimension. This fact, along with the very high levels of intra- and interrater reliability we observed, suggests that the task was significantly easier for listeners than judgments of natural stimuli might have been, even for the EAI task. Second, test stimuli were drawn from the same continuum as were the anchor stimuli. Again, this may have made the task easier than judgments of natural stimuli might be. Finally, listeners were considerably more familiar with the EAI task than with the anchored task. The anchored protocol was described to subjects, but no formal practice was offered. Had subjects been equally familiar with the two tasks, differences between them might have been more pronounced.

Despite these limitations, our results suggest that anchored protocols potentially provide significant improvements in the reliability of perceptual voice evaluations in both clinical and experimental settings. Many questions remain to be addressed before standardized voice assessment protocols may be devised, however. First, what voice qualities actually have perceptual reality (i.e., for which qualities have clinicians developed internal standards)? How many different qualities are uniquely distinguished? What are equal perceptual intervals between the anchors for a given quality? How much do individuals differ in the qualities they distinguish (cf. Bloothooft & Plomp, 1988) and in the nature of their internal standards? As answers to these and similar questions emerge, the reliability of subjective evaluations of voice may approach that of objective measures, thus benefiting diagnosis, treatment, and research practices.

Acknowledgments

We thank our expert listeners Gregory Ator, Evan Bates, Robert Block, Hong-Shik Choi, Pavel Dulgerov, Andrew Erman, Rob Mickel, Douglas Ross, Steven Sloan, Julie Trautmann, Quang Tran, and Mary Walsh for volunteering their time. This research was supported in part by NIDCD grant # DC 00855-01 and funds from the Department of Veterans Affairs.

References

Allen, J., Hunnicutt, M. S., & Klatt, D. (1987). *From text to speech: The MITalk system.* Cambridge: Cambridge University Press.

Berliner, J. E., Braida, L. D., & Durlach, N. I. (1977). Intensity perception. VII. Further data on roving-level discrimination and the resolution and bias edge effects. *Journal of the Acoustical Society of America, 61,* 1256–1267.

Berliner, J. E., Durlach, N. I., & Braida, L. D. (1978). Intensity perception. IX. Effect of a fixed standard on resolution in identification. *Journal of the Acoustical Society of America, 64,* 687–689.

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

Braida, L. D., & Durlach, N. I. (1972). Intensity perception. II. Resolution in one-interval paradigms. *Journal of the Acoustical Society of America, 51,* 483–502.

Braida, L. D., Lim, J. S., Berliner, J. E., Durlach, N. I., Rabinowitz, W. M., & Purks, S. R. (1984). Intensity perception. XIII. Perceptual anchor model of context-coding. *Journal of the Acoustical Society of America, 76,* 722–731.

Cullinan, W. L., Prather, E. M., & Williams, D. E. (1963). Comparison of procedures for scaling severity of stuttering. *Journal of Speech and Hearing Research, 6,* 187–194.

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

Emanuel, F., & Smith, W. (1974). Pitch effects on vowel roughness and spectral noise. *Journal of Phonetics, 2,* 247–253.

Fukazawa, T., & El-Assuooty, A. (1988). A new index for evaluation of the turbulent noise in pathological voice. *Journal of the Acoustical Society of America, 83,* 1189–1193.

Gerratt, B. R., Till, J. A., Rosenbek, J. C., Wertz, R. T., & Boysen, A. E. (1991). Use and perceived value of perceptual and instrumental measures in dysarthria management. In C. A. Moore, K. M. Yorkston, & D. R. Beukelman (Eds.), *Dysarthria and apraxia of speech: Perspective on management* (pp. 77–93). Baltimore: Brookes.

Hillenbrand, J. (1988). Perception of aperiodicities in synthetically generated voices. *Journal of the Acoustical Society of America, 83,* 2361–2371.

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

Kearns, K., & Simmons, N. (1988). Interobserver reliability and perceptual ratings: More than meets the ear. *Journal of Speech and Hearing Research, 31,* 131–136.

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America, 87,* 820–857.

Kojima, H., Gould, W., Lambiase, A., & Isshiki, N. (1980). Computer analysis of hoarseness. *Acta Oto-laryngologica, 89,* 547–554.

Kreiman, J., Gerratt, B., Kempster, G., Erman, A., & Berke, G. (1993). Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research. *Journal of Speech and Hearing Research, 36,* 21–40.

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., Gerratt, B. R., Precoda, K., & Berke, G. S. (1992). Individual differences in voice quality perception. *Journal of Speech and Hearing Research, 35,* 512–520.

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.

Ludlow, C. (1981). Research needs for the assessment of phonatory function. *ASHA Reports, 11,* 3–8. Rockville, MD: American Speech-Language-Hearing Association.

Moll, K. L. (1964). 'Objective' measures of nasality. *The Cleft Palate Journal, 1,* 371–374.

Pollack, I. (1953). The information of elementary auditory displays. II. *Journal of the Acoustical Society of America, 25,* 765–769.

Rossi, M., Pavlovic, C., & Espesser, R. (1990, November). *Reducing context effects in the subjective evaluation of speech quality.* Paper presented at the 120th Meeting of the Acoustical Society of America, San Diego, CA.

Schiavetti, N., Metz, D. E., & Sitler, R. W. (1981). Construct validity of direct magnitude estimation and interval scaling of speech intelligibility: Evidence from a study of the hearing impaired. *Journal of Speech and Hearing Research, 24,* 441–445.

Shrout, P., & Fleiss, J. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420–428.

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

Weismer, G., & Liss, J. (1991). Reductionism is a dead-end in speech research: Perspectives on a new direction. In C. Moore, K. Yorkston, & D. Beukelman (Eds.), *Dysarthria and apraxia of speech: Perspectives on management* (pp. 15–27). Baltimore: Brookes.

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.

Received March 3, 1992
Accepted July 10, 1992

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