Journal of Speech and Hearing Research, Volume 38, 26–32, February 1995

Comparing Reliability of Perceptual Ratings of Roughness and Acoustic Measures of Jitter

C. Rose Rabinov
Jody Kreiman
Bruce R. Gerratt
Steven Bielamowicz
Division of Head and Neck Surgery
UCLA School of Medicine,
and
Audiology and Speech Pathology
VA Medical Center, West Los Angeles

Acoustic analysis is often favored over perceptual evaluation of voice because it is considered objective, and thus reliable. However, recent studies suggest this traditional bias is unwarranted. This study examined the relative reliability of human listeners and automatic systems for measuring perturbation in the evaluation of pathologic voices. Ten experienced listeners rated the roughness of 50 voice samples (ranging from normal to severely disordered) on a 75 mm visual analog scale. Rating reliability within and across listeners was compared to the reliability of jitter measures produced by several voice analysis systems (CSpeech, SoundScope, CSL, and an interactive hand-marking system). Results showed that overall listeners agreed as well or better than "objective" algorithms. Further, listeners disagreed in predictable ways, whereas automatic algorithms differed in seemingly random fashions. Finally, listener reliability increased with severity of pathology; objective methods quickly broke down as severity increased. These findings suggest that listeners and analysis packages differ greatly in their measurement characteristics. Acoustic measures may have advantages over perceptual measures for discriminating among essentially normal voices; however, reliability is not a good reason for preferring acoustic measures of perturbation to perceptual measures.

KEY WORDS: perturbation, acoustic analysis, perception, voice quality

The relationship between acoustic and perceptual measures of voice and their relative importance have long occupied voice researchers. Some authors have argued that perceptual measures of voice have greater content validity, and that acoustic measures are useful only to the extent that they capture perceptual, aerodynamic, or physiological information (e.g., Catford, 1977; Hammarberg, Fritzell, Gauffin, Sundberg, & Weden, 1980; Moll, 1964). Despite such arguments, some clinicians and researchers favor acoustic analysis over perceptual evaluation of pathologic voice. Studies in this tradition emphasize problems of listener unreliability (e.g., Cullinan, Prather, & Williams, 1963) and the lack of standardized terminology (e.g., Jensen, 1965). In contrast, acoustic measures are considered to be well-defined, objective, and reliable. Thus, much effort has been devoted to developing acoustic measures (see Baken, 1987, for review) and to the search for acoustic correlates of various perceptual qualities or pathological states (e.g., Arends, Povel, van Os, & Speth, 1990; Deal & Emanuel, 1978; Hecker & Kreul, 1971; Imaizumi, 1986; Wendahl, 1966; Wendler, Rauhut, & Kruger, 1986; Wolfe & Steinfatt, 1987). Such approaches apparently assume that some day acoustic measures may function in the place of perceptual assessment, thus alleviating concerns about listener unreliability.

However, recent studies suggest that this traditional bias in favor of acoustic analyses of voice quality may be unwarranted. Perceptual data (Gerratt, Kreiman, Antonanzas-Barroso, & Berke, 1993; Kreiman, Gerratt, Kempster, Erman, & Berke, 1993) indicate that some of the noise in listeners' ratings is in fact predictable, and

0022-4685/95/3801-0026

thus potentially controllable. Further, a study comparing several systems for perturbation measurement (Bielamowicz, Kreiman, Gerratt, Dauer, & Berke, 1993) suggested that the reliability of measures produced by different systems may be worse than assumed. Bielamowicz et al. compared measures of signal perturbation produced by CSpeech (ver. 3.1; Paul Milenkovic, Madison, WI), Kay CSL (Kay Elemetrics, Pine Brook, NJ), SoundScope (ver. 1.09; GW Instruments, Cambridge, MA), and by an interactive hand-marking system developed at the VA Medical Center, West Los Angeles, for 50 voices ranging from normal to severely pathologic. In that study, the same digitized voice samples were imported into each system, and values of jitter, shimmer, and the signal-to-noise ratio were calculated (see Appendix for details).

Results for jitter are summarized in Figure 1. Analysis packages varied in their level of overall reliability, with Pearson's r for pairs of comparable algorithms ranging from .20 to .66. However, even systems whose jitter measurements were moderately correlated did not necessarily produce the same numbers for a given voice. None of the lines in Figure 1 has a slope near 1, indicating that all packages under- or over-estimated jitter relative to the others.

Two questions emerge from these findings. First, how do perceptual ratings of voice quality actually compare to acoustic measurements in reliability? Second, how similar are perceptual and acoustic analyses in their characteristics as measurement systems? That is, how similar are the patterns of reliability and agreement among raters to those among analysis systems?

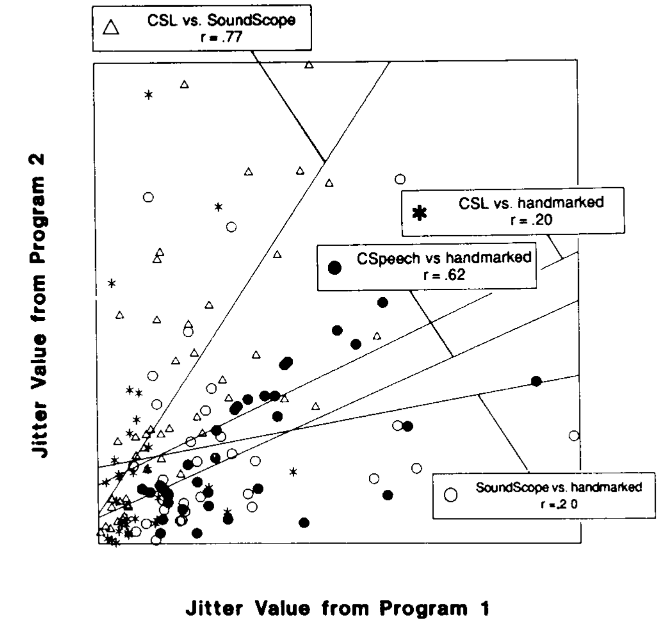

FIGURE 1. Comparison of jitter values produced by four analysis systems. CSpeech values are mean jitter (in msec); values for CSL and SoundScope are percent jitter. The hand-marking system produces values of both mean and percent jitter. Axes in this figure have been scaled from the minimum to the maximum value observed for that pair of algorithms, and one extreme outlier has been deleted from the data for CSL.

To address these questions, we asked listeners to rate the roughness of the voice samples examined by Bielamowicz et al. (1993) and compared these perceptual ratings to the jitter values provided by the various acoustic analysis programs. We chose jitter because it is commonly used as an acoustic measure of vocal quality, because it has been extensively studied, and because measures of jitter are widely available in commercial analysis systems. Further, the different measures of jitter produced by these systems can be easily compared, whereas it is not obvious how to convert the many variants of shimmer derived from different mathematical formulas. Similarly, different algorithms for measuring the harmonics-to-noise ratio (also commonly available in commercial packages) are neither equivalent nor easily transformed into comparable forms (see Bielamowicz et al., 1993, for more discussion). We asked listeners to rate roughness because this scale is also commonly used by clinicians and researchers, and because it is the quality most frequently associated with jitter in the voice literature (e.g., Hillenbrand, 1988; Wendahl, 1966; see Gerratt and Kreiman, in press, for review). Note, however, that our concern is not with the relationship between ratings and measurements per se. Instead, we compared similarities and differences in patterns of reliability among listeners and among analysis systems.

On the surface, comparing reliability of these two analysis methods may appear unreasonable, in that perceptual ratings were made by listeners using a single scale, whereas jitter values were derived from different algorithms. However, listeners apply different criteria when making perceptual judgments, resulting in reduced listener reliability (Kreiman, Gerratt, & Berke, in press; Kreiman, Gerratt, Precoda, & Berke, 1992). These different (but unknown) perceptual algorithms may be considered analogous to the different (but known) algorithms for quantifying perturbation.

Method

Listeners

Ten experienced listeners (otolaryngologists, speech-language pathologists, and phoneticians, including the first three authors) participated in this experiment. Each had a minimum of 2 years experience evaluating pathologic voice quality.

Stimuli

Fifty voices (29 male and 21 female) were selected from an existing library of samples. These voices were also used in the study by Bielamowicz et al. (1993) described above. Voices ranged from normal to severely disordered, with approximately the same number of voices at each of five severity levels from normal to severely dysphonic (as judged by unanimous vote of the authors).

Voice samples were originally obtained by asking speakers to sustain the vowel /a/. Signals were low-pass filtered at 8 kHz, and a 2-second sample was digitized at 20 kHz from the middle of each utterance. Prior to the listening tests, digitized segments were normalized for peak voltage, and

TABLE 1. Number of voices analyzed by each system.

	Severity Level				
	1	2	3	4	5
System					
CSL	10	10	10	10	8
CSpeech	10	10	10	10	8
Hand-marking	10	8	10	4	1
SoundScope	10	9	10	10	9

onsets and offsets were smoothed by 50 msec ramps to eliminate click artifacts.

Procedure

Listeners rated each voice twice, although they were not informed that any voices were repeated. Stimuli were re-randomized for each listener and were presented at a comfortable listening level in free field.

Listeners were tested individually in a sound-treated booth. They were asked to rate the roughness of each voice sample on a 7.5 cm visual analog (VA) scale, using whatever criteria for roughness they normally applied. The scale was displayed horizontally on a computer monitor, and had a resolution of 1 mm. Endpoints were labeled "not rough at all" and "extremely rough." Ten practice trials preceded the experimental session to familiarize listeners with the task.

Acoustic Analyses

Acoustic analysis procedures are described in detail in Bielamowicz et al. (1993). Briefly, voice samples were edited so that each comprised 180 to 200 cycles. Analyses always included this entire segment. For the hand-marking system, either peaks or zero crossings were marked, at the operator's discretion; cycle marking was verified by hand, and obvious errors were corrected. Signals lacking an event that repeated reliably throughout the entire sample were eliminated from statistical analyses. CSpeech, CSL, and SoundScope use fixed F0 tracking strategies, and do not permit access to information about placement of cycle boundaries. Each algorithm did reject some signals, presumably due to F0 tracking errors. The number of voices analyzed by each method is given by severity level in Table 1. See the Appendix for more details.

Results

Intrarater Reliability

All listeners used the entire rating scale when judging vocal roughness, and ratings were fairly evenly distributed across the scale. Levels of test-retest reliability were acceptable for all listeners. Across listeners the correlation (Pearson's r) between the first and second ratings ranged from .75 to .90, with a mean of .83 (SD = .06). On the average, the first and second ratings differed by 9.8 mm (SD = 9.04).

Matched sample *t*-tests compared the first and second ratings of each voice, and indicated that ratings drifted significantly within a listening session. On the average, voices sounded significantly rougher at the second presentation than at the first ($t = -7.56$, df = 499, $p < .01$ one-tailed). Differences between the first and second ratings were also significant for 5 of the 10 individual raters ($p < .01$, adjusted for multiple comparisons).

Of course, computer-based algorithms will always produce identical results under identical conditions. However, changes in analysis parameters within a given system did produce differences in results (Bielamowicz et al., 1993). Repeated analyses using the interactive hand-marking system showed moderate correlations between independent sets of measurements (mean jitter: $r = .81$; percent jitter: $r = .88$; $p < .01$). (One voice that was originally analyzed was rejected as too aperiodic to mark in the second analysis.) Mean jitter values produced by tokenized and untokenized analyses in CSpeech (see Appendix) were correlated at .80; CSL analyses with tolerances of 1 and 20 msec were correlated at .47.

Interrater Reliability

Pairs of raters varied considerably in the extent to which their ratings agreed. Interrater reliability (as measured by Pearson's r) ranged from .32 to .90, with a mean of .71 (SD = .14), compared to a range of .21 to .77 for the different analysis systems (Figure 1). The intraclass correlation (ICC) was calculated using a mixed model analysis of variance (ANOVA) treating voices and listeners as random effects and presentations (first vs. second) as a fixed effect (model [2,1]; e.g., Ebel, 1951; Shrout & Fleiss, 1979). This statistic reflects the overall cohesiveness of a group of raters, as compared to the pairwise comparisons above, and reflects the extent to which the present data might generalize to a new random sample of listeners. For the present data, the ICC = .64, consistent with the variability seen in the pairwise comparisons. Confidence intervals about the ICC were calculated using the formula in Shrout and Fleiss (1979). With 95% certainty, the true ICC value fell in the range $.54 < \rho < .75$.

Examination of patterns of reliability among pairs of raters suggested that subjects fell into two distinct subgroups. One included 7 raters; the other included 3. Within a subgroup, Pearson's r for pairs of raters ranged from .61 to .90, with a mean of .81 (SD = .07); across subgroups, r ranged from .32 to .81, with a mean of .60 (SD = .12). A one-way ANOVA examined the effect of subgroup membership on interrater reliability. This analysis showed that pairs of raters drawn from a single subgroup agreed significantly better than pairs drawn from different subgroups [$F(1,43) = 52.30$, $p < .01$].

Ratings of Individual Voices

Figure 2 shows the width of the 95% confidence interval (in millimeters) about the mean rating of each voice, plotted against the mean rating for that voice. The curve represents the best-fitting quadratic function, as estimated with the Systat nonlinear regression module ($r^2 = .45$; $p < .01$). The

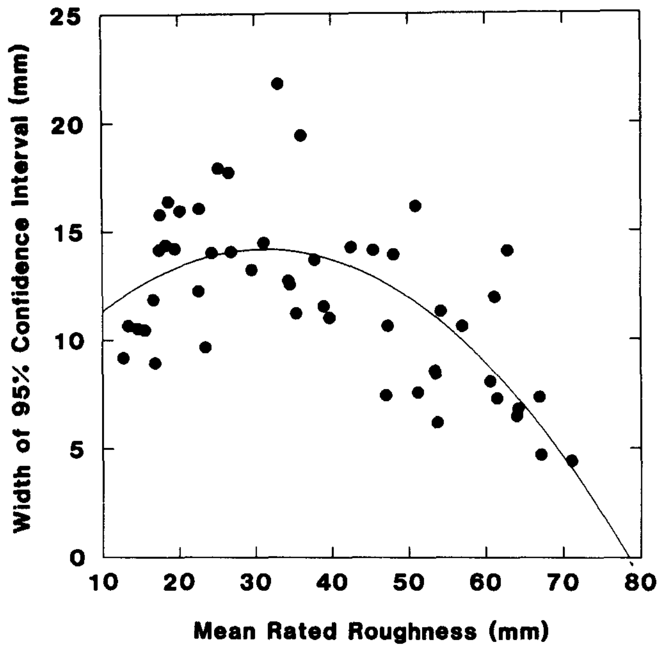

FIGURE 2. Variability in roughness ratings as a function of the mean rating.

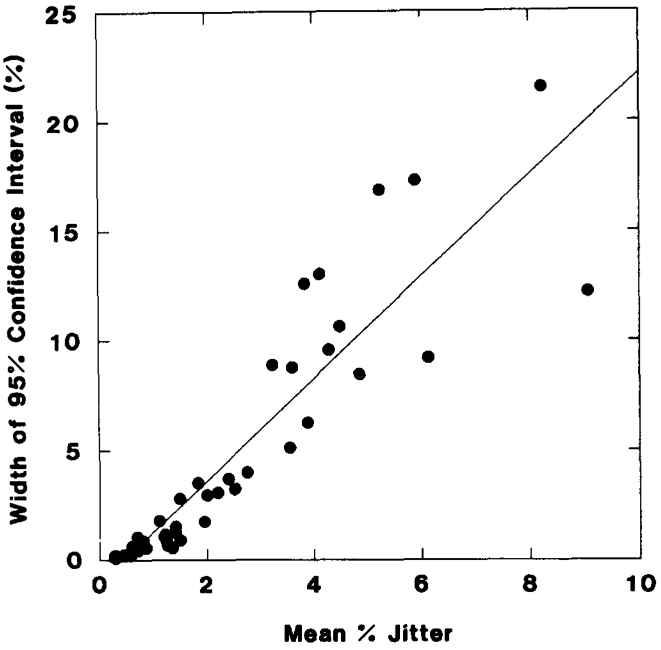

FIGURE 3. Variability in measured jitter as a function of the mean of values produced by four analysis systems.

better the agreement among raters, the smaller the confidence interval. This figure shows the typical pattern (cf. Kreiman et al., 1993) of better agreement among raters (i.e., narrower confidence intervals) for voices at scale extremes, and worse agreement for voices with moderate pathology.[1] The width of the confidence intervals ranged from 4.4 mm (± 2.2 mm) to 21.8 mm (± 10.9 mm) for the 75 mm scale used here.

In contrast, Figure 3 shows the 95% confidence intervals (in percent) around the mean of the percent jitter values produced by the different acoustic analysis systems. (Mean jitter values produced by CSpeech were converted to percent jitter for this analysis.) Confidence intervals in this figure are not directly comparable to those in Figure 2, due to differences in units; however, the *patterns* of changes in measurement reliability with increasing deviation from normal can be compared, and differ markedly. Uncertainty about measured jitter (indicated by larger confidence intervals) increased as a linear function of the mean value [simple linear regression; F(1,47) = 335.35, $p < .01$; $r^2 = .88$]. Confidence interval width ranged from 0.08% to more than 44%.

Figures 4 and 5 show how measurement uncertainty varied with severity of (perceived) pathology, for listeners and analysis systems, respectively.[2] For listeners, the range of variability in ratings increased slightly for voices with moderate pathology, again consistent with previous studies (Kreiman et al., 1993). That is, uncertainty about reliability is greatest for voices in the mid-range of pathology, and least for voices with mild or extremely severe pathology. In contrast, the "variability of the variability" associated with measured jitter increases for the analysis packages once severity exceeds 1, although packages apparently agreed about some voices at all severity levels.

Figure 6 shows the confidence intervals around mean voice ratings, plotted against the confidence intervals for

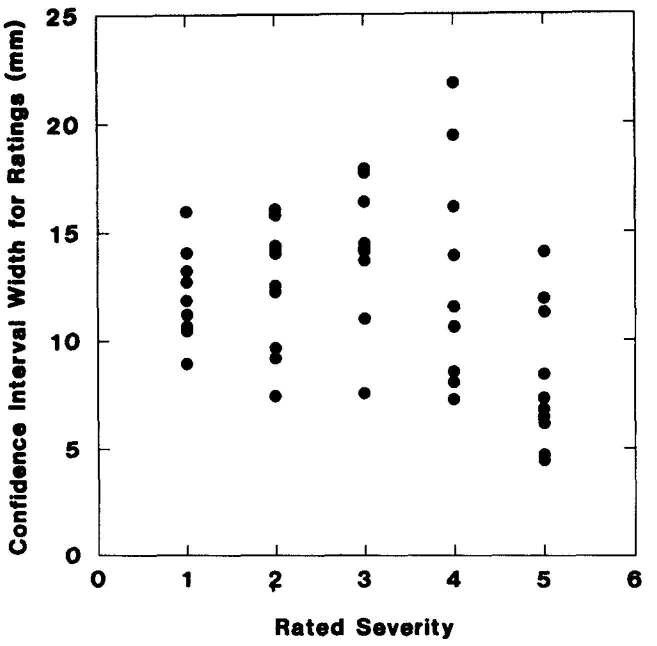

FIGURE 4. Variability in roughness ratings as a function of severity of pathology.

[1] Note that the present stimulus set did not include any normal voices, compared to the set used in Kreiman et al. (1993), which included eight normal samples. Thus, this figure is "shifted to the right" about 10 mm along the x axis as compared to the VA scaling data in our previous study. This shift accounts for the difference in curve shapes in the two studies.

[2] Note that severity is not particularly well correlated with roughness in the present data. Across listeners, Pearson's r ranged from .44 to .87 (mean = .72).

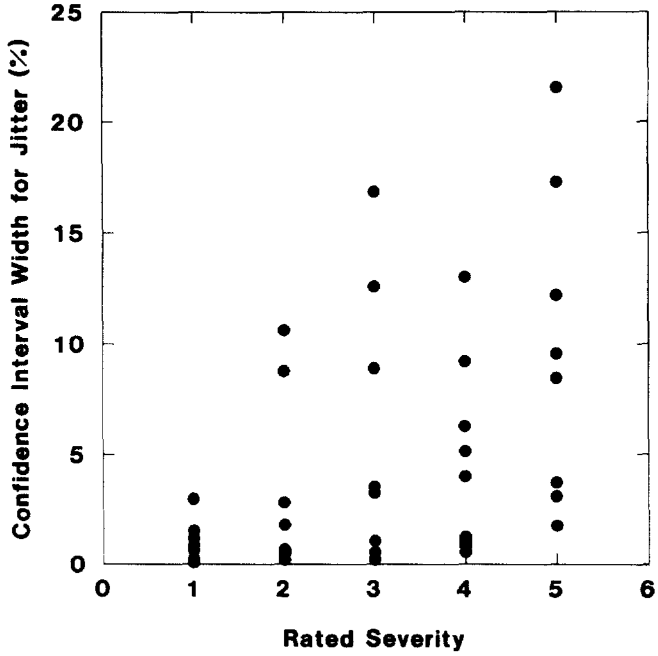

FIGURE 5. Variability in measured jitter as a function of severity of pathology.

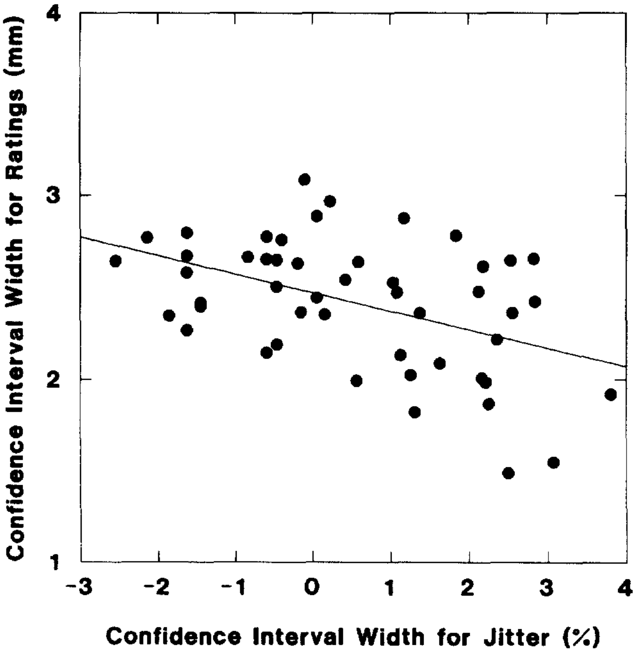

FIGURE 6. Variability in perceived roughness versus variability in measured jitter. Data on both axes have been log-transformed.

mean jitter values. This figure indicates that listeners tend to be most reliable when acoustic measures are most unreliable, and vice versa [simple linear regression; F(1,47) = 12.23, $p < .01$; $r^2 = .21$].

Discussion

Levels of intrarater reliability in this study compare well to those in the literature (e.g., Gerratt et al., 1993; Kreiman et al., 1993), and represent good performance by experienced listeners. At least some test-retest unreliability is caused by systematic drift in ratings. Drift has been reported previously for equal-appearing interval scales (Kreiman et al., 1993), but not for VA scales. Several differences in experimental design between this and the previous study may account for this discrepancy, including use of a shorter VA scale (75 mm vs. 100 mm) with less measurement resolution (.5 mm vs. 1 mm), differences in the distribution of voices across severity levels, and differences in experience levels among listeners. In any event, such drift may be controllable by "anchored" paradigms using fixed comparison stimuli, as we have recently proposed (Gerratt et al., 1993).

In one sense, intrasystem reliability is not a serious issue for acoustic analyses, because computer-based algorithms will always produce identical results under identical conditions. However, changes in analysis parameters within a given system did produce differences in results (Bielamowicz et al., 1993). Recall that the correlation between listeners' first and second ratings of the voices ranged from .75 to .90. The correlation for analyses with different parameters within a given package ranged from .47 to .80. Thus, across voices, performance for even the worst listeners compared well with that of the most consistent jitter analysis systems.

Across pairs of listeners, interrater reliability (measured by Pearson's r) ranged from .32 to .90; the ICC was .64. This compares well to Pearson's r for pairs of analysis systems, which ranged from .21 to .77. However, reliability levels among listeners improved greatly (*r* = .61 to .90) when listeners were compared only to others drawn from the same subgroup. The finding that listeners agreed and disagreed in groups is consistent with multidimensional scaling studies of roughness (Kreiman et al., in press) that reported consistent differences in the strategies listeners used when judging roughness.[3] That study further demonstrated that differences in how listeners focus their attention on the different aspects of multidimensional perceptual qualities are significant predictors of interrater reliability in voice quality ratings. Thus, much of the variation in ratings within and across listeners may not in fact be noise, but may reflect the operation of consistent, predictable perceptual processes. A better understanding of these processes may lead to rating protocols that further enhance listener reliability.

Some disagreement among analysis packages may be attributed to differences in algorithms (see Appendix). However, faulty analysis of noisy or highly perturbed signals also contributed. Because jitter algorithms assume periodicity (i.e., they attempt to locate and measure periodic events in the signal), as the signal departs from this assumption, jitter becomes undefined and values will become increasingly variable. Commercial algorithms make it unfortunately easy to generate perturbation measures when assumptions of periodicity are violated, and make it difficult to determine when violations occur. Users of commercial signal analysis systems must be aware of differences in algorithms, and must be critical in screening voices for periodicity prior to

[3] In particular, listeners varied in how they responded to breathy turbulent noise and tremor.

analysis. Better documentation and education may help reduce the incidence of unsophisticated application of inappropriate analyses.

Interestingly, listeners and acoustic analysis algorithms differed in their properties as measurement systems. As Figure 6 demonstrated, variability in jitter measures increased as rating reliability improved. Variability in rating reliability remained fairly constant across levels of severity, whereas variability in the reliability of measured jitter increased dramatically for voices whose severity exceeded a rating of 1. These results suggest that acoustic measures of jitter have advantages over perceptual measures of roughness for discriminating among essentially normal voices. That is, measures of jitter may be able to resolve very small aperiodicities that are presumably near or below listeners' perceptual thresholds. However, these advantages disappear once signals become even slightly irregular. We therefore question the notion that acoustic measures that depend on the assumption of periodicity (as jitter does) can reasonably substitute for perceptual evaluation in the assessment of pathological vocal quality. (See Gerratt & Kreiman, in press, for further discussion of the question of measurement utility in voice analysis.)

Our results suggest that measured jitter is a function of both signals and algorithms, much as perceptual measures are a function of both signals and listeners. Although standardization of analysis techniques would help solve the problem of disagreements among systems, a standard protocol will still represent a mapping between signals and measured values. The critical issue then becomes defining the "correct" algorithm, the choice of which must depend not only on technical considerations, but also on the purpose for which these measures are intended. As long as acoustic measures are used to detect or define pathology, to aid in diagnosis, to measure the extent of pathology, or to monitor treatment, they must reflect listeners' perceptions reasonably well. Standardization without attention to the characteristics of the application will result in measurements that are not useful.

In conclusion, listeners and analysis packages differ greatly in their measurement characteristics, but reliability is not a good reason for preferring acoustic to perceptual measures. Patterns of unreliability suggest that, for clinical purposes, perceptual measures are probably superior to current acoustic analysis systems, at least for perturbation-based measures. Standardization of acoustic measurement procedures may improve agreement among algorithms, but will not in itself resolve the problem of measurement utility: Standardization without careful attention to all elements in the "speech chain" will not be fruitful.

Acknowledgments

This research was supported in part by NIDCD grant # DC 01797 and by VA research funds. A preliminary version was presented at the Workshop on Standardization in Acoustic Voice Analysis, which took place at the Denver Center for the Performing Arts in February, 1994. We thank Lorraine Ramig, James Hillenbrand, and two anonymous reviewers for their many helpful comments.

References

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

Baken, R. J. (1987). *Clinical measurement of speech and voice.* Boston: College Hill.

Bielamowicz, S., Kreiman, J., Gerratt, B. R., Dauer, M. S., & Berke, G. S. (1993). *A comparison of voice analysis systems for perturbation measurement.* Paper presented at the 125th Meeting of the Acoustical Society of America, Ottawa.

Catford, J. C. (1977). *Fundamental problems in phonetics.* Bloomington: Indiana University Press.

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.

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

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

Gerratt, B. R., & Kreiman, J. (in press). Utility of acoustic measures of voice. In D. Wong (Ed.), *Proceedings of the Workshop on Standardization in Acoustic Voice Analysis.* Denver, CO: Denver Center for the Performing Arts.

Gerratt, B. R., Kreiman, J., Antonanzas-Barroso, N., & Berke, G. S. (1993). Comparing internal and external standards in voice quality judgments. *Journal of Speech and Hearing Research, 36,* 14–20.

Hammarberg, B., Fritzell, B., Gauffin, J., Sundberg, J., & Weden, L. (1980). Perceptual and acoustic correlates of abnormal voice quality. *Acta Otolaryngologica, 90,* 441–451.

Hecker, M. H. L., & Kreul, E. J. (1971). Descriptions of the speech of patients with cancer of the vocal folds. Part I: Measures of fundamental frequency. *Journal of the Acoustical Society of America, 49,* 1275–1282.

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

Horii, Y. (1980). Vocal shimmer in sustained phonation. *Journal of Speech and Hearing Research, 23,* 202–209.

Imaizumi, S. (1986). Acoustic measures of roughness in pathological voice. *Journal of Phonetics, 14,* 457–462.

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.

Koike, Y. (1973). Application of some acoustic measures for the evaluation of laryngeal dysfunction. *Studia Phonologica, 7,* 17–23.

Kreiman, J., Gerratt, B. R., & Berke, G. S. (in press). The multidimensional nature of pathologic vocal quality. *Journal of the Acoustical Society of America.*

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

Milenkovic, P. (1987). Least mean square measures of voice perturbation. *Journal of Speech and Hearing Research, 30,* 529–538.

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

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 Otolaryngologica (Stockholm), Suppl. 338,* 2–24.

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

Wendahl, R. (1966). Laryngeal analog synthesis of jitter and shimmer: Auditory parameters of harshness. *Folia Phoniatrica, 18,* 98–108.

Wendler, J., Rauhut, A., & Kruger, H. (1986). Classification of voice qualities. *Journal of Phonetics, 14,* 483–488.

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

Received March 17, 1994
Accepted June 30, 1994

Contact author: Jody Kreiman, PhD, Audiology and Speech (126), West Los Angeles VA Medical Center, 11301 Wilshire Boulevard, Los Angeles, CA 90073. E-mail: ianpjek@mvs.oac.ucla.edu

Appendix

Summary of Acoustic Analysis Procedures

Information about the algorithms employed by commercial signal analysis packages has been derived from software documentation, literature cited in that documentation, and through discussion with product designers when possible. See Bielamowicz et al. (1993) for more information about these algorithms.

Pitch Tracking

CSL and CSpeech use autocorrelation techniques to track F0. CSL additionally identifies a zero crossing immediately before the "most significant" impulse. The interactive hand-marking system maximizes the Pearson correlation for a moving window around the user-supplied estimate of F0, beginning at a reliably repeating event (peak or zero crossing) identified by the user. SoundScope uses a peak picking strategy, in which individual cycles are identified by marking the zero crossing immediately before the largest peak.

Jitter Algorithms

CSL and SoundScope produce values of Relative Average Perturbation (RAP; Koike, 1973; Takahashi & Koike, 1975), in percent. CSL interpolates when calculating jitter values; SoundScope does not, but offers the option of sampling at 80 kHz when signals are sampled directly. CSpeech calculates mean jitter in milliseconds using a least mean square approach, as described by Milenkovic (1987). The user must supply an estimate of F0. Parabolic interpolation is applied. The hand-marking system calculates mean and percent jitter as described by Horii (1980); linear interpolation is applied when a zero crossing is marked, and parabolic interpolation is used when a peak is chosen (Titze, Horii, & Scherer, 1987).

Analysis Options

CSpeech offers the option of analyzing signals as a whole, or of dividing them into a user-specified number of 100 msec tokens. These tokens may overlap or leave gaps, depending on sample duration and the number of tokens specified. In the present study, intra-program reliability was estimated by comparing untokenized analyses to analyses using 10 tokens.

CSL allows for variation in the amount of perturbation the system will tolerate in calculating perturbation measures. The default is 1 msec of variability in period from cycle to cycle. If a signal varies more, the program will not return a jitter value. Other values may be chosen; for example, Bielamowicz et al. (1993) included tolerances of .5 msec and 20 msec.