Journal of Speech and Hearing Research, Volume 39, 126–134, February 1996

Comparison of Voice Analysis Systems for Perturbation Measurement

Steven Bielamowicz*
Jody Kreiman
Bruce R. Gerratt
Marc S. Dauer**
Gerald S. Berke
Division of Head and Neck Surgery
UCLA School of Medicine
Los Angeles, CA

Dysphonic voices are often analyzed using automated voice analysis software. However, the reliability of acoustic measures obtained from these programs remains unknown, particularly when they are applied to pathological voices. This study compared perturbation measures from CSpeech, Computerized Speech Laboratory, SoundScope, and a hand marking voice analysis system. Sustained vowels from 29 male and 21 female speakers with mild to severe dysphonia were digitized, and fundamental frequency (F_0), jitter, shimmer, and harmonics- or signal-to-noise ratios were computed. Commercially available acoustical analysis programs agreed well, but not perfectly, in their measures of F_0. Measures of perturbation in the various analysis packages use different algorithms, provide results in different units, and often yield values for voices that violate the assumption of quasi-periodicity. As a result, poor rank order correlations between programs using similar measures of perturbation were noted. Because measures of aperiodicity apparently cannot be reliably applied to voices that are even mildly aperiodic, we question their utility in quantifying vocal quality, especially in pathological voices.

KEY WORDS: perturbation, jitter, shimmer, harmonics-to-noise ratio, acoustic analysis

Measures of vocal perturbation have received much attention in the voice literature and are available in most commercial acoustic analysis packages. Such measures are commonly applied in clinical voice analyses, either to document a patient's vocal condition or to track improvements with treatment. They also appear frequently in the experimental literature—for example, in studies of the acoustic correlates of various vocal qualities (e.g., Arends, Povel, van Os, & Speth, 1990; Eskenazi, Childers, & Hicks, 1990; Feijoo & Hernandez, 1990; Hillenbrand, 1988).

Technical issues in perturbation measurement have become prominent in the literature in recent years, as many difficulties in measurement have become apparent (e.g., Cox, Ito, & Morrison, 1989; Doherty & Shipp, 1988; Karnell, 1991; Titze & Liang, 1993; Titze & Winholtz, 1993). Commercially available analysis systems provide concerned users with the image of standardized, well-designed measurement protocols with an acceptably low incidence of technical problems. These systems give users the ability to generate voice analyses with ease and confidence. Use of similar measurement labels for the output (e.g., "mean jitter," "percent shimmer") suggests that results from different programs are comparable. Further, product documentation often makes it difficult to learn how a particular system actually produces its measurements. Little formal information is available about the actual comparability of measures from different analysis packages. Despite this, results

*Currently affiliated with the National Institute on Deafness and Other Communication Disorders, Bethesda, MD
**Currently affiliated with the Department of Radiology, University of Southern California School of Medicine, Los Angeles

0022-4685/96/3901-0126

from such systems are apparently acceptable to speech journal reviewers and editors (e.g., LaBlance & Maves, 1992; Linville & Korabic, 1987; Nittrouer, McGowan, Milenkovic, & Beehler, 1990; Wolfe, Fitch, & Cornell, 1995).

In this study, we tested the extent to which different analysis packages produce comparable measures of perturbation. Three commercially available automated acoustical analysis systems were studied, along with an interactive hand marking system. A similar study (Karnell, Scherer, & Fischer, 1991) used tape-recorded voice samples from two normal speakers to compare perturbation analysis systems in two laboratories and Visipitch measures of perturbation. For the laboratory data, Pearson correlations ranged from .79 to .88 for measures of jitter, with slightly lower correlations among measures of shimmer (r = .60–.80). The Visipitch measures did not agree well with data from either laboratory. Differences in shimmer were attributed to the different analog to digital converters used in the two laboratories and to noise contributed by tape recorders and amplifiers.

The present study used 50 digitized natural signals to eliminate the contaminating effects of analog hardware. We examined human rather than synthetic test signals primarily because pathological voices vary in a potentially great number of ways that interact with perturbation in an unknown manner. Although synthesis of pathological voices has improved in recent years (e.g., Childers & Lee, 1991; Hillenbrand, 1988), both synthesizers and acoustic models remain limited in their ability to accommodate these voices. Thus, it is not currently possible to generate synthetic stimuli that match the rich variation found in a reasonably large sample of natural voices.

Because synthesis of pathological voices inevitably involves inaccurate modeling, testing computational algorithms for perturbation using synthetic signals may falsely lead to conclusions of accuracy and validity, especially for more severely deviant samples. This is unfortunate, because the use of synthetic signals would allow formal evaluation of the accuracy and validity of perturbation measures. These issues cannot be examined directly in the present study. Although the behavior of the various perturbation analysis systems may provide clues about measurement validity, such evidence is unavoidably indirect.

Method

Stimuli

Fifty voices (29 male and 21 female) were selected from an existing library of more than 1,000 samples. To permit examination of the effects of severity of pathology on measurement reliability, voices were chosen so that they ranged from mildly to severely dysphonic. Voices were otherwise unselected with respect to quality. Candidate samples were rated by three experts on a 5-point severity scale, with 1 representing mild dysphonia and 5 representing severe dysphonia. Raters included the second and third authors; each rater had a minimum of 6 years experience evaluating pathological voice quality. Only voices for which all three raters agreed unanimously on a rating were included. Screening continued until 10 voices had been selected at each severity level. Thirty-eight signals were near-periodic (type 1); 4 were type 2, with strong subharmonics and/or modulations; and 8 included segments of apparently chaotic vocal fold vibration (type 3) (see Titze, 1995, for more discussion of signal typing in perturbation analysis).

Recording Procedures

Each voice was originally recorded by asking speakers to sustain the vowel /ɑ/ with comfortable F_0 and loudness. A head-mounted microphone (Sony ECM-79B) was placed 10 cm from the subject's lips. Signals were low-pass filtered at 8 kHz and directly digitized at 20 kHz, using a 12 bit analog-to-digital converter. One hundred eighty to 200 contiguous cycles of phonation were selected from the middle of each utterance for acoustic analysis.

Acoustic Analyses

Three commercially available programs were evaluated: CSpeech (ver. 4.0; Paul Milenkovic, Madison, WI), Computerized Speech Laboratory (CSL; Kay Elemetrics, Pine Brook, NJ), and SoundScope (ver. 1.09; GW Instruments, Cambridge, MA). An interactive hand marking program developed at the VA West Los Angeles Voice Laboratory was also evaluated. All of these programs permit importation of previously digitized signals, thus avoiding the confounding effects of tape wow and flutter or noise introduced by second generation digitization of analog recordings (Doherty & Shipp, 1988; Titze, Horii, & Scherer, 1987). Other systems requiring analog input (e.g., Visipitch; Kay Elemetrics, Pine Brook, NJ) were excluded.

Information about the algorithms employed by commercial signal analysis packages is often unfortunately difficult to obtain. For this reason, we describe the systems used here at some length. Our presentation has been derived from software documentation, literature cited in that documentation, and through discussion with product designers when possible.

F_0 tracking. CSL and CSpeech use autocorrelation techniques to track F_0. CSL additionally identifies a zero crossing immediately before the "most significant" impulse. The hand marking system maximizes the Pearson correlation for a moving window around a user-supplied estimate of F_0, 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 (Gold, 1962).

Jitter algorithms. CSL and SoundScope produce values of Relative Average Perturbation (RAP; Koike, 1973; Takahashi & Koike, 1975), in percent. RAP is defined as:

$$\text{RAP} = \frac{\sum_{i=1}^{n} |\text{Period Difference}_i|}{\sum_{i=1}^{n} \text{Period}_i} \qquad (1)$$

where P_i is the period of the i^{th} cycle, n is the number of consecutive cycles analyzed, and:

$$\text{Period Difference}_i = \frac{P_{i-1}+P_i + P_{i+1}}{3} - P_i \qquad (2)$$

where P_{i-1} is the period of the $i - 1^{th}$ cycle and P_{i+1} is the period of the $i + 1^{th}$ cycle. CSL interpolates when calculating jitter values; SoundScope does not, but offers the option of sampling at 80 kHz when signals are digitized directly.

CSpeech calculates mean jitter in ms using a least mean square approach to estimate the differences in duration of two consecutive periods, given an estimate of F_0 supplied by the user (Milenkovic, 1987):

$$\text{Jitter} = |t_p(n_0) - t_p(n_0 - N_p)| \qquad (3)$$

where t_p is the period, n_0 is the reference sample position within the speech waveform, and N_p is the number of samples contained in a period. Parabolic interpolation is applied.

The hand marking system reports both mean and percent jitter as described by Horii (1980). Mean jitter is defined as:

$$\text{Mean Jitter} = \frac{\sum_{i=1}^{n-1} |P_i - P_{i+1}|}{n - 1} \qquad (4)$$

Percent jitter is calculated by dividing mean jitter in ms by the mean period in ms and multiplying by 100. Linear interpolation is applied when a zero crossing is marked, and parabolic interpolation is used when a peak is chosen (Titze et al., 1987).

Shimmer algorithms. The hand marking system calculates mean shimmer in dB as described by Horii (1980):

$$\text{Mean Shimmer} = 20/(n - 1) \sum_{i=1}^{n-1} |\log_{10}(A_i/A_{i+1})| \qquad (5)$$

where A_i is the peak amplitude of the i^{th} cycle, A_{i+1} is the peak amplitude of the $i + 1^{th}$ cycle, and n is the number of contiguous cycles analyzed. Percent shimmer is also provided by the hand marking system as:

$$\text{Percent Shimmer} = \frac{100 * \text{mean shimmer}}{(20/n)\sum_{i=1}^{n} \log_{10} A_i} \qquad (6)$$

SoundScope reports the amplitude perturbation quotient (APQ; Takahashi & Koike, 1975). If A_i and A_{i+1} are the peak amplitudes of two consecutive cycles and:

$$\text{A differences}_i = \frac{\left[\sum_{k=-5}^{k=5} A_{i+k}\right]}{11} - A_i \qquad (7)$$

then:

$$\text{Shimmer(\%)} = \frac{100 * \text{A differences}_i}{\left[\sum_{K=-5}^{K=5}\right] \Big/ 11} \qquad (8)$$

CSpeech reports percent shimmer as:

$$\text{Shimmer} = |100(1 - K)| \qquad (9)$$

where

$$K = R+(1+R^2)^{1/2} \qquad (10)$$

and where:

$$R = \frac{\dfrac{s(s)^T}{s(s_d)^T} - \dfrac{s_d(s_d)^T}{s(s_d)^T}}{2} \qquad (11)$$

where s and s_d are element vectors described as:

$$s = [s(nT)] \quad \text{and} \quad s_d = [s(nT - t_p)] \qquad (12)$$

where n is the integer and T is the interval at which the signals s(t) and $s(t - t_p)$ are sampled.

CSL's shimmer measure is based on RAP. The sum of the absolute sample values is calculated on a period by period basis for the entire sample, and RAP is calculated using formulae (1) and (2) above (Davis, 1981), substituting the amplitude of the i^{th} cycle (A_i) for the period of the i^{th} cycle.

Harmonics-to-noise ratio. The hand marking system measures HNR in dB as the ratio of harmonic energy, H, to noise energy, N (Yumoto, Gould, & Baer, 1982):

$$\text{HNR} = \text{H/N} \qquad (13)$$

The harmonic component of the averaged acoustic wave, $f_A(\tau)$, is given as:

$$H = n\int_{0}^{T} f_A{}^2(\tau)d\tau \qquad (14)$$

where τ ranges over the duration of a cycle, T_i is the duration of the i^{th} period in a window of n cycles, T is the maximum of T_i in the window, and n is the number of cycles measured.

The noise component of the i^{th} period is equal to $f_i(\tau) - f_a(\tau)$, where τ ranges from 0 to T_i and is described by:

$$N = \sum_{i=1}^{n} \int_{0}^{T} [(f_i(\tau) - f_A(\tau)]^2 \, d\tau \qquad (15)$$

In SoundScope, a cycle P is chosen and the average waveform in 31 cycles centered at P is calculated. The difference of the i^{th} period centered about P and the harmonic wave is calculated, yielding a noise signal. With M equal to the number of points in one cycle, N equal to the number of cycles averaged to estimate the harmonic component, $\text{Wave}_{\text{time}}$ equal to voltage as a function of time, and mod M equal to the integer remainder, HNR is then calculated as:

$$\text{HNR} = 10 \log_{10}(H_i/N_i) \qquad (16)$$

where:

$$H_i = \sum_{n=1}^{m} \text{Wave}_{avgi}{}^2(n) \qquad (17)$$

$$N_i = \sum_{n=1}^{m} [\text{Wave}_{\text{avgi}}(n) - \text{Wave}_{\text{time}}(n)]^2 \qquad (18)$$

and

$$\text{Wave}_{\text{avgi}}(N) = 1/N \sum_{i=N/2}^{N/2-1} \text{Wave}_{\text{time}}((iM + N) \bmod M) \qquad (19)$$

CSL also follows Yumoto et al.'s (1982) description for HNR calculation. The program selects 8 consecutive cycles with period variation less than 6.25% and then computes HNR for these cycles.

The signal-to-noise ratio (SNR) is calculated in CSpeech as described by Milenkovic (1987):

$$\text{SNR} = 10 \log_{10}[s(s^T/E_{\text{opt}})] \qquad (20)$$

in which $s * s^T$ gives the energy in the acoustic signal within the observation interval and E_{opt} is the interpolated minimum value of E_0. E_0 is given in the formula:

$$E_0 = [1/(1 + K^2)][-2Ks(s_d)^T + ss^T + K^2 s_d(s_d)^T] \qquad (21)$$

where the variables are described in formulae 10–12.

Analysis options. CSpeech offers the option of analyzing signals as a whole, or of dividing them into a user-specified number of 100 ms tokens. These tokens may overlap or leave gaps, depending on sample duration and the number of tokens specified.

CSL allows for variation in the amount of perturbation the system will tolerate in calculating perturbation measures. The default is 1 ms of variability in period from cycle to cycle, but other values may be specified. If deviation for a given cycle exceeds the specified amount, that cycle is omitted from the analysis; if the number of cycles rejected exceeds the limit, no value of jitter is returned (although shimmer is estimated in most cases). Thus, analyses with different perturbation tolerances may include different numbers of cycles for a given signal.

Analyses performed. Measures of F_0, jitter, shimmer, and the signal- or harmonics-to-noise ratio were generated using each system. Fundamental frequency was estimated for CSpeech and hand marking analyses by averaging the frequency of 10 contiguous cycles of phonation from the middle of the segment. CSpeech, CSL, and SoundScope use fixed F_0 tracking strategies and do not permit access to information about placement of cycle boundaries. In the hand marking system, 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.

All 50 voices were independently reanalyzed with the hand marking system to assess the effect of differences in event selection on perturbation measures. The effect of tokenizing (1 vs. 10 tokens) in CSpeech and changes in perturbation tolerance (1 ms vs. 20 ms) in CSL were also assessed with separate analyses of the complete voice sets.[1] In addition, within each analysis for the commercial systems, 13 voice signals (selected at random) were reanalyzed to check for operator error. Values were also recalculated for all outliers in plots comparing the various systems.

Results

Voices Analyzed

The number of voices analyzed by each method is given by severity level in Table 1. The hand marking system rejected many more voices than did the other systems, including 5/38 type 1 signals and all type 2 and type 3 signals. Independent operators agreed perfectly about which voices were not analyzable. The waveform in Figure 1 exemplifies the difficulties encountered in tracking F_0 in the rejected samples. The hand marking system attempted to track the first major negative peak in this waveform; parabolic interpolation was applied.[2] Crosses in the figure indicate suggested locations of cycle boundaries. Many of the peaks are irregularly shaped, making it difficult to determine precisely where to place a mark for a given peak. For example, should the third cross be placed earlier? Further, large changes in waveform shape occur throughout this segment, with peaks appearing, disappearing, and changing in amplitude. It appears at first glance as if the second and subsequent crosses are mislocated. However, a small pos-

[1] Analyses were also attempted with a tolerance of 5 ms. However, all voices in the set were rejected at this level.

[2] Similar difficulties emerged when positive peaks or zero crossings were tracked in this waveform, although they are not apparent in the segment presented here.

TABLE 1. Number of voices analyzed by each system.

	Severity level				
Algorithm	**1**	**2**	**3**	**4**	**5**
CSL (1 ms tolerance)					
F_0	10	10	10	10	10
jitter	10	10	10	10	8
shimmer	10	10	10	10	10
SNR	10	10	10	10	10
CSL (20 ms tolerance)					
All measures	10	10	10	10	10
CSpeech (1/10 tokens)					
All measures	10	10	10	10	8
Hand marking					
F_0	10	10	10	7	5
jitter	10	8	10	4	1
shimmer	10	8	10	4	1
SNR	10	8	10	4	1
SoundScope					
F_0	10	10	10	10	10
jitter	10	9	10	10	9
shimmer	9	9	10	10	5
SNR	8	9	8	9	4

itive peak occurs between the arrows and crosses in the first two cycles. If the second cross is moved earlier in the cycle (to the second arrow in the figure), then the first cross should also be moved to the location indicated by the first arrow, to reflect the position of this small peak. If the first cross is also moved, however, we are no longer marking the first major negative peak. In this signal, there is no way to mark cycles that is consistent with respect to both wave shape and peak locations. Such paradoxes are common in the analysis of dysphonic voices, even for voices that appear highly periodic. Such paradoxes led the hand marking system to reject signals as unanalyzable. The commercial analysis systems all returned a full set of perturbation parameters for this sample.

Intraprogram Reliability

TABLE 2. Intraprogram reliability: Pearson's *r* for values obtained with different settings within a program.

Algorithm	Jitter	Shimmer	HNR/SNR
CSL (1 vs. 20 ms tolerance)	.47	.67	.98
CSpeech (1 vs. 10 tokens)	.80	.92	.97
Hand marking (independent analyses by two operators)			
M	.81	.96	.87
percent	.88	.95	

In one sense, intraprogram reliability is a trivial issue, because computer-based algorithms will always produce identical results under identical conditions. However, the analysis options described above did affect measured perturbation. For each system, Table 2 lists the Pearson's correlation between measures of jitter, shimmer, and HNR/SNR produced with the different analysis settings.

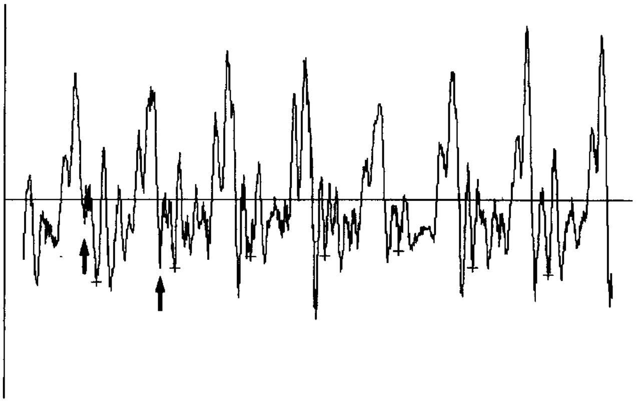

FIGURE 1. Excerpt from an acoustic waveform that was rejected as unanalyzable by the hand marking system, but was evaluated by all three commercial analysis systems.

As might be expected, analysis settings differed in their effects on perturbation measurements. CSL analyses using different tolerance values included different numbers of cycles and would not be expected to agree well, particularly for severely pathological voices. Tokenizing (CSpeech) is averaging across a series of windows; short-term variation in jitter levels would affect the series of means. For the hand marking system, correlations reflect differences in starting point (Jafari, Till, Truesdell, & Law-Till, 1993) and differences in choice of a peak or zero (Deem, Manning, Knack, & Matesich, 1989; Titze & Liang, 1993), both of which affect perturbation measures. Jitter reflects these effects more than shimmer, because jitter is particularly sensitive to time-varying event selection (e.g., Deem et al., 1989; Titze & Liang, 1993).

Interprogram Reliability

Means, ranges, and standard deviations for the different perturbation measures are given for each analysis system in Tables 3–6. Values in these tables are collapsed across severity levels (hence the large standard deviations). Cases where data are missing for one or more systems have been deleted.[3]

For each measure, interprogram reliability was assessed by examining rank-order correlations (Spearman's rho) between values produced by all pairs of comparable algorithms. Thus, programs measuring mean jitter were com-

TABLE 3. F_0 values (Hz) produced by each program.

	Package			
Statistic	**CSpeech**	**CSL**	**SoundScope**	**Hand marking**
min	82	83	85	83
max	262	273	274	273
M	162.3	164.3	162.9	164.3
SD	49.9	51.2	49.0	51.3

Note. N = 42.

[3] Thus no type 2 or type 3 signals are included in these results, because the hand marking system rejected all such voices as unmarkable.

TABLE 4. Jitter measures produced by each program.

	Mean jitter (ms)		Percent jitter		
	CSpeech	Hand marking	CSL	SoundScope	Hand marking
min	0.013	0.023	0.212	0.120	0.314
max	0.080	0.138	4.964	2.292	2.963
M	0.037	0.054	1.175	0.704	0.837
SD	0.019	0.026	1.058	0.570	0.585
N	33	33	31	31	31

pared only to others measuring mean jitter, percent shimmer to percent shimmer, and so on.[4] Nonparametric comparisons were used because of the limited range of values observed for some measures, because of the small number of observations in some cells, and because the existence of extreme outliers violated the assumptions necessary for parametric comparisons. Note, however, that a high rank-order correlation may mask significant differences among programs in the actual values produced, as discussed below.

In addition, we examined the relationship between reliability and severity by computing separate correlations for different severity levels. In these analyses, data from adjacent severity levels were combined (levels 1 and 2; 2 and 3; 3 and 4; and 3, 4, and 5) to reduce variability due to the limited number of observations in some cells. Results are shown in Figures 2–5.

The programs examined here produced grossly comparable measures of F_0, although values produced by the hand marking system were significantly greater than those produced by CSpeech (matched sample *t*-test; $t\ (41) = -3.37$, $p < .01$).[5] For the complete voice set, Spearman's rho for pairs of algorithms ranged from .97 to .996, as shown on the left side of Figure 2. Correlations decreased significantly with severity of pathology [simple linear regression; $F\ (1, 22) = 9.38$, $p < .01$, $r^2 = .30$].

Agreement among programs varied much more for jitter (Figure 3). The hand marking system produced significantly larger values of mean jitter than did CSpeech [matched sample *t*-test; $t\ (32) = -4.61$, $p < .01$], and CSL produced consistently larger values of percent jitter than SoundScope did [matched sample *t*-test; $t\ (46) = 4.66$, $p < .01$]. For the complete voice set, Spearman's rho for pairs of comparable programs ranged from .33 (CSL vs. hand marking system) to .80 (SoundScope vs. CSL). Overall, the hand marking system agreed best with CSpeech (rho = .61), but quite poorly with the other programs (rho = .33 and .34). Figure 3 shows no consistent relationship between interprogram reliability and increasing dysphonia [simple linear regression; $F\ (1, 14) = 0.75$, $p > .01$].

Significant mean differences were found between the percent shimmer values produced by the algorithms studied here (matched sample *t*-tests; CSpeech vs. hand marking: $t\ (32) = 6.18$, $p < .01$; SoundScope vs. hand marking: $t\ (30) = 7.20$, $p < .01$; CSpeech vs. SoundScope: $t\ (41) = 3.68$, $p < .01$). Spearman's rho for pairs of algorithms ranged from .81 to .89 (Figure 4). Rank-order correlations for values of mean shimmer (CSL and the hand marking system) ranged from rho = .82 to .87 across severity levels. Correlations among measures of percent shimmer varied more across severity levels (rho = .66 to .95), again with no systematic change with severity [$F\ (1, 14) = 3.67$, $p > .01$].

Finally, matched sample *t*-tests showed significant mean differences in HNR/SNR values produced by all pairs of packages except CSpeech and the hand marking system, which did not differ [CSpeech vs. CSL: $t\ (47) = 13.03$, $p < .01$; CSpeech vs. SoundScope: $t\ (37) = 5.63$, $p < .01$; CSL vs. SoundScope: $t\ (37) = -7.52$, $p < .01$; CSL vs. hand marking: $t\ (32) = -11.64$, $p < .01$; SoundScope vs. hand marking: $t\ (27) = -6.73$, $p < .01$). Spearman correlations for the HNR/SNR measures produced by the different packages are shown in Figure 5. Correlations across severity levels

[4] In theory, mean jitter values could be converted to percent jitter for purposes of these comparisons by dividing by the mean F_0. However, significant differences in the F_0 values found by different algorithms (as described below) suggest that this conversion would introduce additional error into jitter measures. For this reason, the original measurement units have been retained.

[5] *P* values have been adjusted for multiple comparisons throughout this study.

TABLE 5. Shimmer values produced by each program.

	Mean shimmer (dB)		Percent shimmer		
	CSL	Hand marking	CSpeech	SoundScope	Hand marking
min	0.185	0.143	1.270	1.482	0.690
max	0.763	0.788	10.300	10.100	4.397
M	0.414	0.386	3.949	3.520	1.989
SD	0.168	0.165	2.472	1.697	0.868
N	33	33	31	31	31

TABLE 6. Harmonics- or signal-to-noise ratio values produced by each program.

	CSpeech	CSL	SoundScope	Hand marking
min	13.53	−1.67	7.88	7.42
max	26.37	17.83	21.78	35.09
M	18.48	8.58	14.34	21.31
SD	3.43	6.15	3.40	6.39

Note. N = 28.

ranged from .23 to .86; correlations *increased* reliably with increasing severity [F (1, 22) = 8.03, $p < .01$; $r^2 = .27$].

Discussion

Reliability of F_0 measurement decreased significantly with increasing dysphonia, and the number of voices for which cycle boundaries could not be consistently identified increased for some analysis systems. In addition, significant differences were found between the F_0 values produced by CSpeech and the hand marking system. As discussed in the introduction, the validity of perturbation measurements cannot be examined directly when natural signals are used as stimuli. However, because the validity of perturbation measures depends on extremely accurate and consistent determination of cycle boundaries, these results imply that such measures may be invalid for voices with even mild dysphonia. The hand marking system allows users to examine placement of cycle boundaries before proceeding with perturbation analyses, so the validity of the analysis can be assessed. This control is not possible with the other systems, which do not permit access to information about boundary placement or adjustment of criteria for accepting or rejecting an analysis.

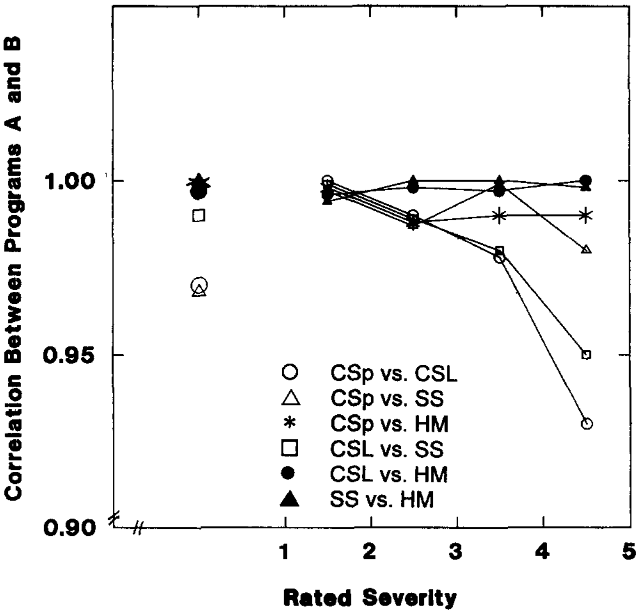

FIGURE 2. Interprogram reliability (as measured by Spearman's rho) for measures of F_0. Larger symbols at the left of the figure represent correlations between pairs of algorithms across the entire voice set. Smaller symbols connected by lines represent correlations at different severity levels. Adjacent severity levels have been combined, as described in the text. CSp = CSpeech; HM = hand marking system; SS = SoundScope.

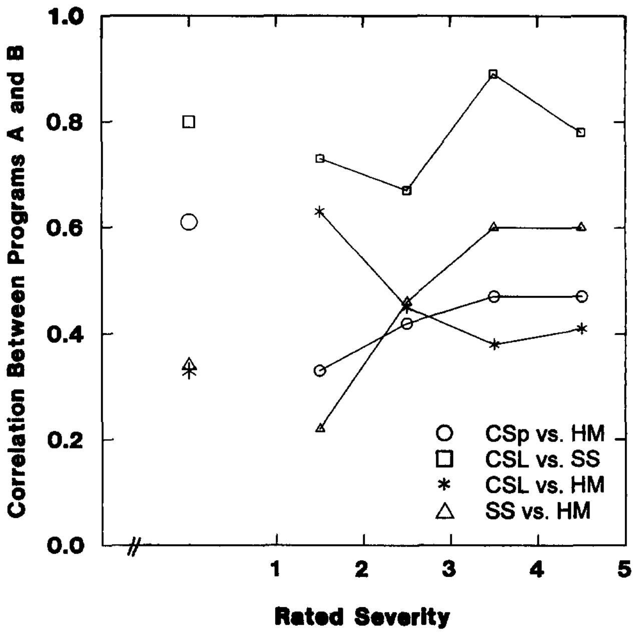

FIGURE 3. Interprogram reliability (as measured by Spearman's rho) for comparable measures of jitter. Larger symbols at the left of the figure represent correlations between pairs of algorithms across the entire voice set. Smaller symbols connected by lines represent correlations at different severity levels. Adjacent severity levels have been combined, as described in the text. CSp = CSpeech; HM = hand marking system; SS = SoundScope.

Shimmer measures varied much more in reliability at all levels of severity than F_0 measures, and reliability was not related significantly to increasing dysphonia. Significant differences were found between values of percent shimmer, but not mean shimmer, probably because of differences in computational formulae. In particular, the hand marking system uses the log of the sum of the amplitudes in the denominator, whereas SoundScope does not include the log function. For moderately to severely pathological samples (severity levels 3, 4, and 5), agreement was better for comparisons involving the hand marking system (which eliminated samples where F_0 tracking was doubtfully valid) than for comparisons not involving the hand marking system, again suggesting that analysis validity is questionable for these voices.

Overall reliability was worse for jitter and HNR than for F_0 and shimmer. For jitter, values for CSL and SoundScope (both of which measure jitter as RAP, in percent) showed the highest rank-order correlations at every level of severity, but values from CSL were significantly larger than those produced by SoundScope. Reliability was poor at all severity levels for other algorithms. Both CSpeech and the hand marking system measure average cycle-to-cycle deviations in period length; however, differences in F_0 tracking, calculation methods, and interpolation algorithms may account for the rather low correlations between these systems (for which significant differences in F_0 were reported above).

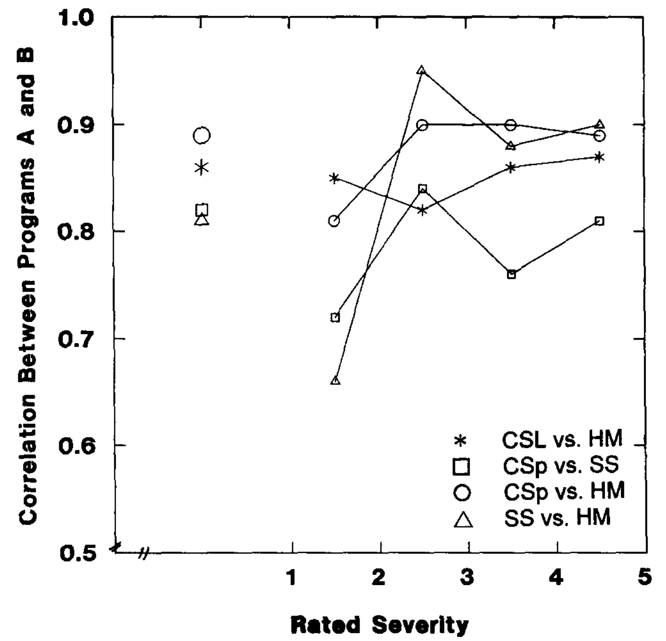

FIGURE 4. Interprogram reliability (as measured by Spearman's rho) for measures of shimmer. Larger symbols at the left of the figure represent correlations between pairs of algorithms across the entire voice set. Smaller symbols connected by lines represent correlations at different severity levels. Adjacent severity levels have been combined, as described in the text. CSp = CSpeech; HM = hand marking system; SS = SoundScope.

Finally, SoundScope measures "percent jitter" as RAP, whereas the hand marking system measures jitter factor. Values for these different algorithms are not well correlated.

HNR/SNR is calculated by the various analysis systems using unique modifications of Yumoto's original algorithm (Yumoto et al., 1982) and by varying the window of averaging. Thus, differences among the values produced are expected, although the particular pattern of results reported here is difficult to explain. Values from the hand marking system are larger than those from CSL and SoundScope, as suggested by the computational formulae above, but do not differ from CSpeech, as might also be predicted. The apparent increase in reliability with severity is probably artifactual: Values for the most severe voices are almost certainly invalid, because of large cycle-to-cycle differences in period length (e.g., Cox et al., 1989). Errors in F_0 tracking would compound this problem, making these measures doubly suspect. Algorithms incorporating dynamic time warping may increase accuracy for such measures (e.g., Qi, 1992).

These findings indicate that differences in F_0 tracking procedures and in perturbation algorithms produce large disagreements among analysis packages in the values produced for measures of perturbation. Differences in F_0 tracking need not be large to produce large effects on perturbation measures, and packages that do not allow access to information about the placement of cycle boundaries make it virtually impossible to determine if perturbation analyses are valid. As mentioned, validity cannot be examined directly with natural stimuli. However, the large differences found in the measurements from the various systems imply questionable accuracy, especially for the more severely pathological samples. Such inaccuracy further suggests basic problems of validity.

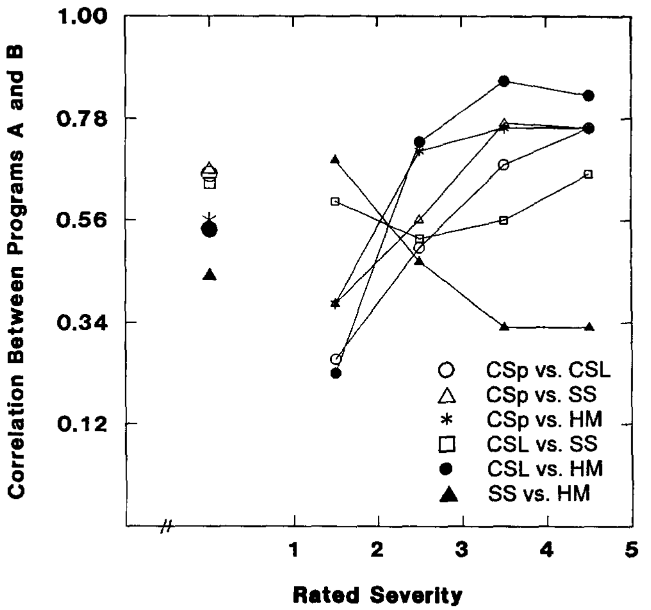

FIGURE 5. Interprogram reliability (as measured by Spearman's rho) for measures of the harmonics- or signal-to-noise ratio. Larger symbols at the left of the figure represent correlations between pairs of algorithms across the entire voice set. Smaller symbols connected by lines represent correlations at different severity levels. Adjacent severity levels have been combined, as described in the text. CSp = CSpeech; HM = hand marking system; SS = SoundScope.

The present results parallel those of Titze and Liang (1993) for normal voices. They reported good reliability for measures of F_0, but poor reliability for jitter, for 20 male and 20 female speakers. Problems with reliability were linked to the specific F_0 tracking strategies applied. Waveform matching methods gave the least jitter of all, followed by zero crossing methods and by peak picking. Standardization of procedures would eliminate this source of noise, but pathological voices often lack a repeating event of a given type. For example, a voice may have a reliable zero crossing but no peak that can be marked, or vice versa. Thus, attempts to standardize procedures will almost certainly increase errors in marking, at least for some voices.

Finally, our results suggest that an important consideration in measurement development and use is the robustness of the measure with respect to aperiodicity—that is, the extent to which it can be accurately applied to aperiodic signals. Mean F_0 and, to a lesser extent, shimmer depend less critically on precise F_0 tracking than do jitter and HNR/SNR. Small differences in the placement of cycle boundaries have little effect on the mean value of F_0, because errors are small relative to the unit of measurement used for mean F_0 (usually 0.1 to 0.01 Hz) and because errors are generally random and thus have little cumulative effect on a mean value. Similarly, very small errors in placement of cycle boundaries have a small effect on shimmer, because measuring shimmer depends primarily on locating the peak amplitude within a given cycle, and marking errors are not generally of sufficient magnitude to eliminate an entire peak from a cycle. Thus, these quantities can be measured with reasonable reliability across severity levels. In contrast, jitter and HNR/SNR critically depend on accurate F_0 tracking; even tiny errors in placing cycle boundaries add noise to these measures. Consequently, as error in cycle matching or in placement of cycle boundaries increases with worsening vocal pathology, jitter validity decreases. The fact that jitter (a measure of aperiodicity) lacks robustness with respect to aperiodicity seems to us a fatal paradox inherent in this measure. Consequently, despite its theoretical and historical appeal, we believe abandonment of jitter as a measure of pathological voice should be seriously considered.

In conclusion, commercially available acoustical analysis programs agree fairly well in their measures of F_0, implying such measures are grossly accurate. Differences in F_0 that did occur cannot be explained, because commercial systems do not allow access to information about placement of cycle boundaries. Measures of perturbation in the various analysis packages use different algorithms, provide measures in different units, and often produce values for signals whose cycle boundaries cannot be accurately determined. As a result, poor correlations between programs using similar measures of perturbation were noted. Differences in shimmer values are probably largely due to differences in algorithms. Differences in values of jitter and HNR/SNR are probably due to differences in F_0 marking and to differences in algorithms. Values for many analyses are probably invalid. Time-warping algorithms may lead to more valid measures of HNR, at least for some voices. However, jitter cannot be measured validly when voices are aperiodic. The fact that measures of aperiodicity apparently cannot be reliably applied to signals that are even slightly aperiodic leads us to question their utility in analyzing vocal quality, especially in pathological voices.

Acknowledgments

We thank Scott Selbie, PhD, for his help determining the formal equivalence of different perturbation algorithms. SoundScope software was donated by GW Instruments. No financial support was provided by any of the organizations whose software was evaluated here. This research was supported in part by National Institute on Deafness and Other Communication Disorders grant DC 01797 and by Veterans Administration Merit Review funds, and was carried out at the VA Medical Center, West Los Angeles. A preliminary version was presented at the 124th Meeting of the Acoustical Society of America in Ottawa, Canada, on May 19, 1993.

References

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

Childers, D. G., & Lee, C. K. (1991). Vocal quality factors: Analysis, synthesis, and perception. *Journal of the Acoustical Society of America, 90,* 2394–2410.

Cox, N. B., Ito, M. R., & Morrison, M. D. (1989). Technical considerations in computation of spectral harmonics-to-noise ratios for sustained vowels. *Journal of Speech and Hearing Research, 32,* 203–218.

Davis, S. B. (1981). Acoustic characteristics of normal and pathological voices. *ASHA Reports* (11), 97–115.

Deem, J. F., Manning, W. H., Knack, J. V., & Matesich, J. S. (1989). The automatic extraction of pitch perturbation using microcomputers: Some methodological considerations. *Journal of Speech and Hearing Research, 32,* 689–697.

Doherty, E. T., & Shipp, T. (1988). Tape recorder effects on jitter and shimmer extraction. *Journal of Speech and Hearing Research, 31,* 485–490.

Eskenazi, L., Childers, D. G., & Hicks, D. M. (1990). Acoustic correlates of vocal quality. *Journal of Speech and Hearing Research, 33,* 298–306.

Feijoo, S., & Hernandez, C. (1990). Short-term stability measures for the evaluation of vocal quality. *Journal of Speech and Hearing Research, 33,* 324–334.

Gold, B. (1962). Computer program for speech extraction. *Journal of the Acoustical Society of America, 34,* 916–921.

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.

Jafari, M., Till, J., Truesdell, L. F., & Law-Till, C. (1993). Time-shift, trial, and gender effects on vocal perturbation measures. *Journal of Voice, 7,* 326–336.

Karnell, M. (1991). Laryngeal perturbation analysis: Minimum length of analysis window. *Journal of Speech and Hearing Research, 34,* 544–548.

Karnell, M. P., Scherer, R., & Fischer, L. B. (1991). Comparison of acoustic voice perturbation measures among three independent voice laboratories. *Journal of Speech and Hearing Research, 34,* 781–790.

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

LaBlance, G. R., & Maves, M. D. (1992). Acoustic characteristics of post-thyroplasty patients. *Otolaryngology—Head & Neck Surgery, 107,* 558–563.

Linville, S. E., & Korabic, E. W. (1987). Fundamental frequency stability characteristics of elderly women's voices. *Journal of the Acoustical Society of America, 81,* 1196–1199.

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

Nittrouer, S., McGowan, R. S., Milenkovic, P., & Beehler, D. (1990). Acoustic measurements of men's and women's voices: A study of context effects and covariation. *Journal of Speech and Hearing Research, 33,* 761–775.

Qi, Y. (1992). Time normalization in voice analysis. *Journal of the Acoustical Society of America, 92,* 2569–2576.

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

Titze, I. R. (1995). *Workshop on Acoustic Voice Analysis: Summary Statement.* Denver: National Center for Voice and Speech.

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

Titze, I. R., & Liang, H. (1993). Comparison of F_0 extraction methods for high-precision voice perturbation measurements. *Journal of Speech and Hearing Research, 36,* 1120–1133.

Titze, I. R., & Winholtz, W. S. (1993). Effect of microphone type and placement on voice perturbation measurements. *Journal of Speech and Hearing Research, 36,* 1177–1190.

Wolfe, V., Fitch, J., & Cornell, R. (1995). Acoustic prediction of severity in commonly occurring voice problems. *Journal of Speech and Hearing Research, 38,* 273–279.

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 November 7, 1994
Accepted July 31, 1994

Contact author: Jody Kreiman, PhD, Division of Head/Neck Surgery, UCLA School of Medicine, 31-24 Rehabilitation Center, Los Angeles, CA 90095-1794. E-mail: jkreiman@ucla.edu