Title
Perception of aperiodicity in pathological voice

Journal
Journal of the Acoustical Society of America, 117(4)

ISSN
0001-4966

Authors
Kreiman, J
Gerratt, B R

Publication Date
2005-04-01

Peer reviewed

Perception of aperiodicity in pathological voice

Jody Kreiman[a] and Bruce R. Gerratt[b]
Division of Head and Neck Surgery, UCLA School of Medicine, 31-24 Rehab Center, Los Angeles, California 90095-1794

(Received 20 December 2003; revised 20 December 2004; accepted 22 December 2004)

Although jitter, shimmer, and noise acoustically characterize all voice signals, their perceptual importance in naturally produced pathological voices has not been established psychoacoustically. To determine the role of these attributes in the perception of vocal quality, listeners were asked to adjust levels of jitter, shimmer, and the noise-to-signal ratio in a speech synthesizer, so that synthetic voices matched naturally produced tokens. Results showed that, although listeners agreed well in their judgments of the noise-to-signal ratio, they did not agree with one another in their chosen settings for jitter and shimmer. Noise-dependent differences in listeners' ability to detect changes in amounts of jitter and shimmer implicate both listener insensitivity and inability to isolate jitter and shimmer as separate dimensions in the overall pattern of aperiodicity in a voice as causes of this poor agreement. These results suggest that jitter and shimmer are not useful as independent indices of perceived vocal quality, apart from their acoustic contributions to the overall pattern of spectrally shaped noise in a voice. [DOI: 10.1121/1.1858351]

PACS numbers: 43.71.Bp, 43.71.Gv [PFA] Pages: 2201–2211

I. INTRODUCTION

Jitter, shimmer, and noise-to-signal ratios (NSRs) are the cornerstones of acoustic measurement of voice signals, and are often regarded as indices of the perceived quality of both normal and pathological voices. In general, applications of acoustic measures to assess vocal quality derive their validity from the relevance of specific acoustic properties of the signal to auditory perceptions of voice. Researchers typically use correlation or regression techniques to demonstrate the extent to which such measures explain or predict listeners' scalar quality judgments. However, observed associations between acoustic and perceptual measures have varied considerably across studies. For example, correlations between measures of jitter and ratings of both breathiness and roughness have ranged from 0 to about 0.7 (for review see Heiberger and Horii, 1982; Kreiman and Gerratt, 2000). Multidimensional scaling studies (which examine the perceptual dimensions that underlie perceived vocal similarity) have also provided inconsistent results with respect to the role that jitter, shimmer, and the NSR play in determining quality (Kreiman *et al.*, 1990; Kreiman and Gerratt, 1996). Such variability in results has undermined confidence in both the acoustic measures and their perceptual interpretation.

Although hundreds of studies describing, evaluating, and applying measures of noise and acoustic signal perturbation have been published (Buder, 2000), the perceptual salience of these attributes remains poorly understood. A discrepancy exists between the results of early synthesis studies and findings from later investigations examining this association in naturally produced voices (Heiberger and Horii, 1982). Synthesis studies (Wendahl, 1963, 1966a, b; Coleman and Wendahl, 1967) used sawtooth waves with added jitter ($\pm$1–50 Hz around a mean $f0$ of 100 or 200 Hz) or shimmer (alternate periods reduced in amplitude by 1–6 dB). Near-perfect correlations were observed between the amount of jitter or shimmer and judgments of relative roughness for these nonspeech stimuli. More recently, Hillenbrand (1988) used synthetic vowels to examine the univariate relationships between jitter, shimmer, and noise and ratings of breathiness and roughness. He reported that roughness ratings increased with levels of jitter and shimmer up to about 2% jitter and 2-dB shimmer, after which increasing jitter and shimmer levels did not increase perceived roughness. Similarly, as the NSR increased, so did breathiness ratings.

Several limitations are inherent in these synthesis studies. The synthesis techniques employed in the earliest studies (Wendahl, 1963, 1966a, b) did not include vocal-tract filtering, and used a highly unnatural sawtooth source along with levels of jitter and shimmer (up to $\pm$50%) that greatly exceed those usually associated with the human voice (e.g., Andrianopoulos *et al.*, 2001; Munoz *et al.*, 2003). Stimuli lacked pitch contours or formant structure, and varied in only one acoustic dimension at a time. Further, a relatively small number of stimuli differed in relatively large acoustic steps, making it easy for listeners to perform the paired comparison task reliably (Heiberger and Horii, 1982). These factors could account for the high correlations with perceived vocal roughness, and limit the extent to which early studies explain how listeners hear naturally produced voice signals. (Hillenbrand, 1988, also commented that his stimuli sounded somewhat unnatural.) Further, previous studies assessing the role of noise and perturbation in determining voice quality have always assessed quality in terms of specific scales like breathiness, roughness, or hoarseness. However, the reliability of such scales has been repeatedly questioned, and their validity as measures of quality is also questionable (e.g., Jensen, 1965; Kreiman, Gerratt, and Berke 1994; see Kreiman *et al.*, 2005, for review).

A final limitation of previous synthesis studies is the fact

[a]Electronic mail: jkreiman@ucla.edu
[b]Electronic mail: bgerratt@ucla.edu

that jitter, shimmer, and noise have been manipulated independently of one another, and the perceptual interactions among these three variables have not been investigated. Acoustically, these attributes are obviously related. For example, adding aspiration noise to a signal adds jitter and shimmer; adding jitter also adds shimmer as changes in period length move harmonics toward and away from vocal-tract resonances; and adding jitter and shimmer produces measurable changes in the NSR (Hillenbrand, 1987). Naturally produced voice signals include all these attributes, and separating them analytically has proven difficult (Hillenbrand, 1987; Fukazawa *et al.*, 1988; de Krom, 1993; Titze, 1995; Michaelis *et al.*, 1997; Murphy, 1999).

In contrast to synthesis studies, correlational investigations of jitter, shimmer, and noise perception in naturally produced voices have produced highly variable results. Many low or nonsignificant correlations have been reported (e.g., Martin *et al.*, 1995; de Krom, 1995; Deal and Emanuel, 1978; Prosek *et al.*, 1987). However, correlational approaches are of limited use in resolving the issues surrounding the perceptual importance of vocal aperiodicity, because they cannot provide evidence about cause and effect. Lack of information about how (or whether) an acoustic variable evokes perception of a particular vocal quality leaves researchers and clinicians to make assumptions without evidence about the validity and utility of such measures. For example, it is impossible to determine if statistically significant changes in some parameter are actually perceptually meaningful or not in the absence of a demonstrated cause and effect relationship between an acoustic variable and a perceptual outcome.

In this study, we used a method of adjustment task to examine the perceptual significance of jitter, shimmer, and the NSR. Listeners were asked to manipulate jitter, shimmer, and NSR levels in high-quality synthetic voice stimuli, until the synthetic voices matched naturally produced target voices. This approach avoids the use of verbal rating scales, because listeners compare the stimuli directly in terms of their overall similarity. Further, in the method of adjustment task, subjects manipulate acoustic variables directly, so the association between signal and percept need not be inferred from correlation. Finally, in this task the perceptual importance of acoustic variables is evaluated in the same complex, multidimensional context in which the variables naturally occur. In particular, this study examined the perceptual interactions of jitter and shimmer with noise. A previous study (Gerratt and Kreiman, 2001) found good agreement among listeners for the NSR. It is possible that, because noise comprises jitter and shimmer acoustically, listeners agree about overall levels of noise, but not about the levels of jitter and shimmer present in a voice. On the other hand, if measures of jitter and shimmer are valid indices of perceived vocal quality, then listeners should agree well in the levels of these variables they select in the method of adjustment task.

II. EXPERIMENT 1

A. Rationale

To examine the contributions of jitter, shimmer, and noise to voice quality, listeners in this experiment were asked to adjust the levels of jitter, shimmer, and/or noise present in a synthetic stimulus until that stimulus matched a naturally produced voice sample as closely as possible. We then examined the extent to which listeners agreed about how much of each variable was necessary for the synthetic and natural voice samples to sound the same, consistent with the ANSI definition of sound quality as ''that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar'' (ANSI Standard S1.1.12.9, p. 45, 1960). If jitter, shimmer, and noise contribute independently to the perceived quality of a voice, then listeners should agree well in the values they choose when making their adjustments.

B. Method

1. Voice samples

The voices of 20 speakers with voice disorders (10 males and 10 females) were selected from a large library of samples recorded under identical conditions as part of a clinical voice evaluation. Selection was random except that samples with period doubling or biphonation were excluded because jitter and shimmer are undefined for such signals (e.g., Titze, 1995). Speakers ranged in age from 22 to 78 years (mean=49.8 years; sd=17.2 years) and represented a variety of primary diagnoses, including mass lesions of the vocal folds (7), glottal incompetence (6), chronic laryngitis (4), adductory spasmodic dysphonia (2), and Parkinson disease (1). Severity of vocal deviation was rated by consensus vote of the authors. Ratings ranged from 2–5 on a 5-point scale (where 1=normal quality), and averaged 3.55 (sd =0.97).

During the voice evaluation, speakers sustained the vowel /a/ as steadily as possible. Voice signals were transduced with a 1-in. Bruel & Kjaer condenser microphone held a constant 5 cm off axis. They were then low-pass filtered at 8 kHz and directly digitized at 20 kHz. A 1-s segment was excerpted from the middle of these productions, antialias filtered, and downsampled to 10 kHz for further analysis.

2. Analysis and synthesis methods

Analysis techniques have been described in detail elsewhere (Gerratt and Kreiman, 2001; Gabelman, 2003). Briefly, formant frequencies and bandwidths were estimated using linear predictive coding analysis with a window of 25.6 ms (increased to 51.2 m when stimulus $f0$ was near or below 100 Hz). The glottal volume velocity derivative was estimated by interactively inverse filtering a single glottal pulse from the microphone recordings, using the method described by Javkin *et al.* (1987). The output of the inverse filter was least-squares fit with a modified Liljencrants–Fant (LF) source model (Fant *et al.*, 1985), and the parameters of the best-fitting LF model were used to specify the harmonic component of the voice source in the synthesizer. The shape of the inharmonic part of the voice source (the noise spectrum) was estimated by applying cepstral comb filtering to remove periodic source components following the method described by de Krom (1993). The residual was then inverse filtered to remove the effects of vocal-tract resonances, leav-

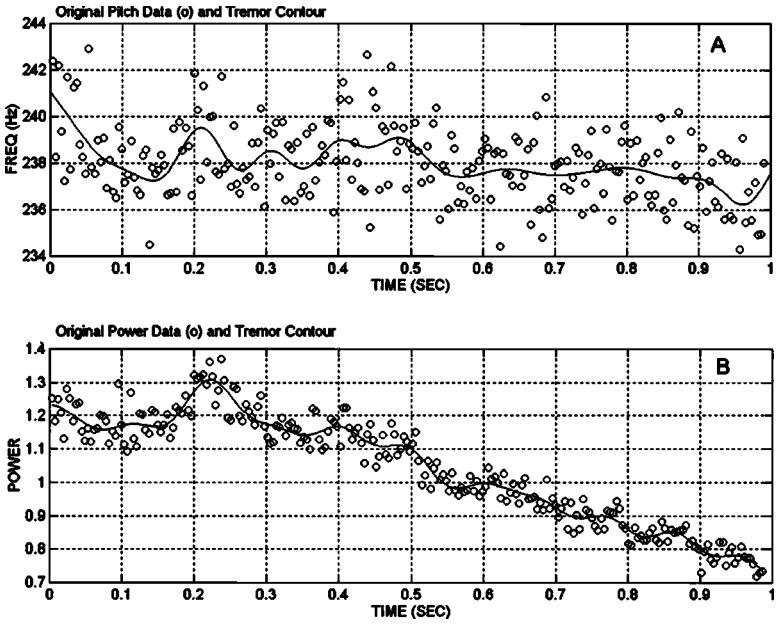

FIG. 1. Modeling of jitter, shimmer, and low-frequency modulations of frequency and power (tremors). (A) Output of $f0$ analysis algorithm. Solid line shows the low-frequency pitch track, and open circles show deviations about this contour (jitter). (B) Output of power analysis algorithm. Solid line shows the low-frequency power track, and open circles show deviations about this contour (shimmer). Units for power are arbitrary.

ing the inharmonic part of the source, which was fitted with a 25-segment piecewise linear approximation. Finally, $f0$ was tracked pulse by pulse on the time-domain waveform by an automatic algorithm. Location of cycle boundaries was verified by the first author. Tremor rates were estimated by visual inspection of smoothed plots of $f0$ versus time (Kreiman *et al.*, 2003). Estimated tremor rates averaged 6.45 Hz, with a range of 2 to 10 Hz and a median of 6 Hz.

Variability in $f0$ was modeled as follows (Fig. 1). The estimated tremor rate served as a dividing line between slow frequency and amplitude modulations (tremors) on the one hand and fast variations (jitter and shimmer) on the other. To model tremor, the $f0$ track was low-pass filtered with the cutoff value equal to the estimated tremor rate, and period lengths were altered to incorporate this slowly varying frequency track [Fig. 1(a)]. Low-frequency amplitude modulations were similarly generated for each voice [Fig. 1(b)], and the power of each cycle was modified to incorporate observed low-frequency modulations in amplitude.[1] Jitter was modeled by altering the duration of each cycle by an amount sampled from a high-pass filtered, normally distributed random sequence whose variance was determined by the desired level of jitter. The filter's cutoff frequency again equaled the estimated tremor value. Shimmer was similarly modeled by altering the power of each cycle of phonation, with power values sampled at random from a normal distribution of values whose variance corresponded to the desired amount of shimmer. (See Gabelman, 2003, for more detail.)

Stimuli were synthesized using a custom formant synthesizer implemented in MATLAB (MathWorks, Natick, MA; Gerratt and Kreiman, 2001).[2] The synthesizer sampling rate was fixed at 10 kHz. The following procedure was applied to overcome quantization limits on modeling $f0$. Source pulses were synthesized pulse by pulse. A single LF-modeled source pulse was upsampled from 10 to 40 kHz and used as a basis for each pulse in the time series. A plot of the desired $f0$ values versus time was generated for the 1-s synthetic sample, taking into account the tremor contour and requested level of jitter. As the series of pulses was constructed, precise ending times for each LF pulse were tracked, and this curve was interpolated to find $f0$ for the next pulse in the series. The basic LF pulse was then time warped (stretched or compressed) to obtain the appropriate period length, and sample points were calculated accordingly. A similar power versus time curve was constructed for the time series, reflecting amplitude tremor and shimmer, and this curve was interpolated to determine the power for each successive cycle in the time series. The overall effect is equivalent to digitizing an analog pulse train with pulses of the exact desired frequencies and amplitudes at the fixed 10-kHz sample rate.

A 100-tap finite impulse response filter was constructed for noise synthesis, with its shape corresponding to the 25-segment piecewise linear approximation fit to the inharmonic part of the voice source derived through comb filtering. A spectrally shaped noise time series was then created by passing white noise through this filter. The jittered, shimmered, and tremored LF pulse train was antialias filtered and downsampled to 10 kHz, and then added to this noise time series to create a complete source time series. The source was filtered through the vocal-tract model to generate a preliminary version of the synthesized voice. Finally, formant frequencies, bandwidths, and LF parameters were adjusted to provide good perceptual matches to the target voices, and then held constant across experimental conditions.

3. Listening pretest

A listening pretest was used to verify the accuracy of the synthesis. Prior to this pretest, estimated values of the NSR, jitter, and shimmer were adjusted in the synthesizer by the first author as necessary to provide the best possible perceptual match to the natural voice samples, because measurement of jitter, shimmer, and the NSR is difficult and often inaccurate when phonation departs from periodicity (e.g., Titze, 1994; Bielamowicz *et al.*, 1996). Synthetic copies of each of the 20 natural voice samples were then generated,

using the methods described above and these perceptually adjusted levels of jitter, shimmer, and noise. Twenty listeners (UCLA students and staff; 20 to 53 years of age; mean age =26.4 years; sd=9.93 years) heard pairs of voices. On half of the trials, a synthetic voice sample was paired with its natural counterpart, and on the other half stimuli were identical. Each pair was repeated 3 times, for a total of 120 trials/listener.

For each trial, listeners were asked to judge whether the two samples were the same or different, and to rate their confidence in their response on a 5-point scale ranging from ''positive'' to ''wild guess.'' Listeners were not allowed to replay the stimuli before responding. Order of voices in ''different'' pairs was randomized, and the stimulus pairs were rerandomized for each listener. Listeners were tested individually in a double-walled sound suite. To approximate normal listening conditions, stimuli were presented in free field at a comfortable constant listening level. Testing lasted approximately 15 min.

To provide a measure of the average discriminability of the synthetic and natural tokens, responses were pooled across listeners. Overall rates of correct and incorrect ''same'' responses (hits and false alarms) were calculated for each voice. Hit rates ranged across voices from 85%–98.3% (mean=91.8%; sd=3.66%); false-alarm rates ranged from 51.7%–85% (mean=65.7%; sd=11.18%). The probability of any correct response (''same'' or ''different'') ranged across voices from 54.2%–70% (mean=63.1%; sd=5.02%).

Same/different responses for each voice were combined with confidence ratings to create a 10-point scale ranging from ''positive voices are the same'' to ''positive voices are different.'' For example, ''same'' responses with confidence equal to 1 (positive) were recoded as ''1,'' ''same'' responses with confidence equal to 5 (wild guess) were recoded as ''5,'' ''different'' responses with confidence equal to 5 were recoded as ''6,'' and ''different'' responses with confidence equal to 1 were recoded as ''10.'' Receiver operating characteristics (ROCs) consisting of 9 points each (10 recoded response categories minus 1) were constructed from these recoded data following the procedure described by Green and Swets (1966; see also MacMillan and Creelman, 1991). The area under the ROC for each voice was calculated, along with 99% confidence intervals around these values. In all cases, these confidence intervals included the chance value of 0.5. These data, combined with consistently high false-alarm rates, indicate that listeners were unable to consistently distinguish the synthetic copies from the natural samples. We conclude that the synthesis technique is able to model the quality of the natural target voice samples adequately for purposes of the following experiments.

4. Listening task

Seventy listeners participated in experiment 1. They ranged in age from 19–57 years (mean age=26.7 years; sd =7.83 years). All reported normal hearing. Listeners were unselected with respect to experience with voice disorders, and most were phonetically- and otolaryngologically naive.

Listeners were tested individually in a sound-treated room. Each listener participated in 20 trials, one for each voice. Voices were presented to each listener in a unique random order, in free field through two loudspeakers at a constant comfortable listening level (to approximate normal listening conditions). At the beginning of each trial, listeners heard the natural voice sample paired with a copy of that voice synthesized without jitter, shimmer, or noise. They were told that one voice was naturally produced, and that the other was a synthetic copy. They were then asked to adjust the quality of the second voice by moving a sliding cursor or cursors on a computer monitor with a mouse, until the two voices matched on the target dimension or dimensions as closely as possible. These cursors controlled the jitter, shimmer, and NSR levels present in the synthetic voice. Depending on condition, listeners adjusted either one, two, or all three parameters in a given trial. Listeners heard a given voice in a single condition only. Which task a listener performed for a given voice was assigned at random, with the constraint that ten listeners performed each of the seven possible tasks (jitter only; shimmer only; noise only; jitter +shimmer; jitter+noise; shimmer+noise; or jitter+shimmer +noise) for each voice. Listeners were encouraged to play the natural and synthetic stimuli as often as they liked in any order, and could make as many adjustments as necessary to achieve the best possible match to the target voice.

Scale displays were 115 mm in length. The NSR scale ranged from −50 to 0 dB (noise-free to extremely noisy); the jitter scale ranged from 0% to 3%; and the shimmer scale ranged from 0 to 2 dB. These values were chosen based on measurements of the test voices, on pilot tests, and on data from Hillenbrand (1988), who reported that increasing jitter beyond 2.5% and shimmer beyond 2 dB had little additional effect on perceived vocal roughness. Ranges of values for jitter, shimmer, and the NSR vary widely across previous studies, for both pathological and normal voices, due to differences in measurement techniques, instrumentation, and computational algorithms. In the present research, stimuli were frequency- and amplitude demodulated prior to perturbation analysis to remove the contributions of tremor from measured values. Values of these variables in the synthesizer were also calculated independently of the effects of tremor. For these reasons, the ranges used here for these variables are lower than those reported by some authors (see Buder, 2000, for review). Parameters not being manipulated in a given trial remained set in their ''off'' position (0% jitter, 0-dB shimmer, −50-dB NSR).

Prior to the experiment, the synthesizer was demonstrated and two practice items were presented, so that listeners could become thoroughly familiar with the task and with the sounds of different amounts of jitter, shimmer, and noise, individually and in combination. Practice continued until subjects were satisfied that they understood and could perform the task; it lasted about 15 min on average. The total test session lasted 1.5–2 h. Listeners were encouraged to take breaks during this period as necessary to maintain attention and motivation.

C. Results

Figure 2 shows the distributions of jitter, shimmer, and NSR responses, summed over voices and experimental con-

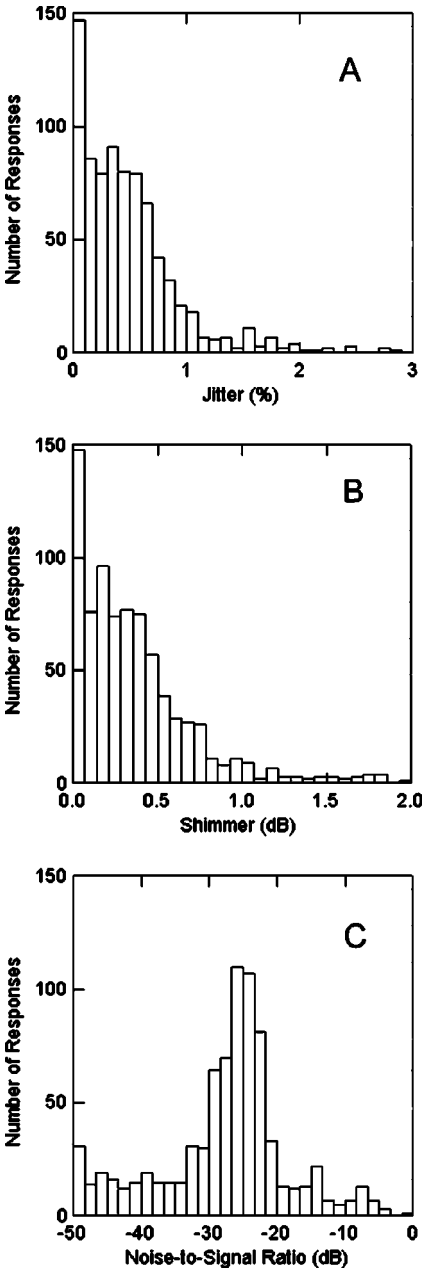

FIG. 2. Distribution of responses for jitter (A), shimmer (B), and the NSR (C), pooled across the 20 stimuli in the method-of-adjustment task in experiment 1.

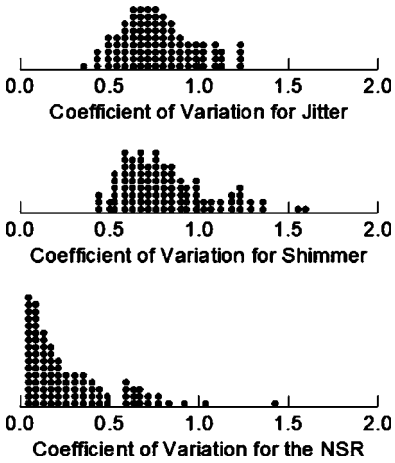

FIG. 3. Distribution of coefficient of variation values for the jitter, shimmer, and NSR responses in experiment 1, across experimental conditions and stimuli.

ditions. Across conditions and voices, listeners used the full range for each of these scales. Responses for jitter and shimmer clustered toward the lower third of each scale, with listeners adding only small amounts of jitter and/or shimmer on most trials. In contrast, NSR responses were approximately normally distributed.

To measure response variability, we calculated the coefficient of variation (the standard deviation divided by the mean) for the jitter, shimmer, and NSR responses for each voice. Figure 3 shows the distribution of variation coefficients for each measure across the 20 stimulus voices. For each of the individual voices, variability of the jitter and shimmer responses exceeded that of the NSR responses. On average, jitter and shimmer responses were more than 5 times more variable than NSR responses, and in one case they were over 20 times more variable. These differences were significant across voices [$F(2,297)=162.12$, $p<0.01$; Bonferroni *post hoc* comparisons $p<0.01$].

Jitter and shimmer responses varied significantly with the particular listening task. Listeners used significantly less jitter and shimmer when asked to adjust all three parameters at once than they did when matching jitter or shimmer alone [jitter: $F(3,796)=4.32$, $p<0.01$; shimmer: $F(3,796)=4.14$, $p<0.01$; Bonferroni *post hoc* comparisons $p<0.01$]. In contrast, listeners always adjusted the NSR to similar levels, even when they were also adding jitter and/or shimmer to the voice, so no significant effect of response condition on NSR values was observed [$F(3,796)=1.01$, n.s.]. However, coefficients of variation did not vary significantly by experimental task [jitter: $F(4,95)=0.288$, $p>0.01$; shimmer: $F(4,95)=1.05$, $p>0.01$; NSR: $F(4,95)=0.525$, $p>0.01$].

Variability in NSR responses *decreased* with increasing rated severity of vocal deviation. In other words, the worse the voice sounded, the better the listeners agreed in their NSR responses ($r=-0.64$, $p<0.01$). No significant relationship between severity of deviation and response variability was observed for jitter or shimmer (jitter: $r=-0.23$, n.s.; shimmer: $r=-0.16$, n.s.). However, variability in jitter and shimmer responses did increase significantly with the NSR (jitter: $r=0.68$, $p<0.01$; shimmer: $r=0.74$, $p<0.01$). When the NSR was low, listeners' responses rarely exceeded 1% jitter or 1-dB shimmer. However, as the NSR increased, so did response variability, with some listeners adding little or no jitter or shimmer to the synthetic stimuli, and others adding large amounts. This contrasts sharply with NSR responses: The higher the NSR, the better listeners agreed in their responses ($r=-0.66$, $p<0.01$).

To examine the relationship between response variability and the spectral shape of the harmonic part of the source (Gerratt and Kreiman, 2001), we calculated the difference in the amplitudes of the first two harmonics ($H1$-$H2$) from FFT spectra of the LF source pulses. This measure provides one index of the source spectral slope, independent of the influence of the vocal-tract transfer function: The larger

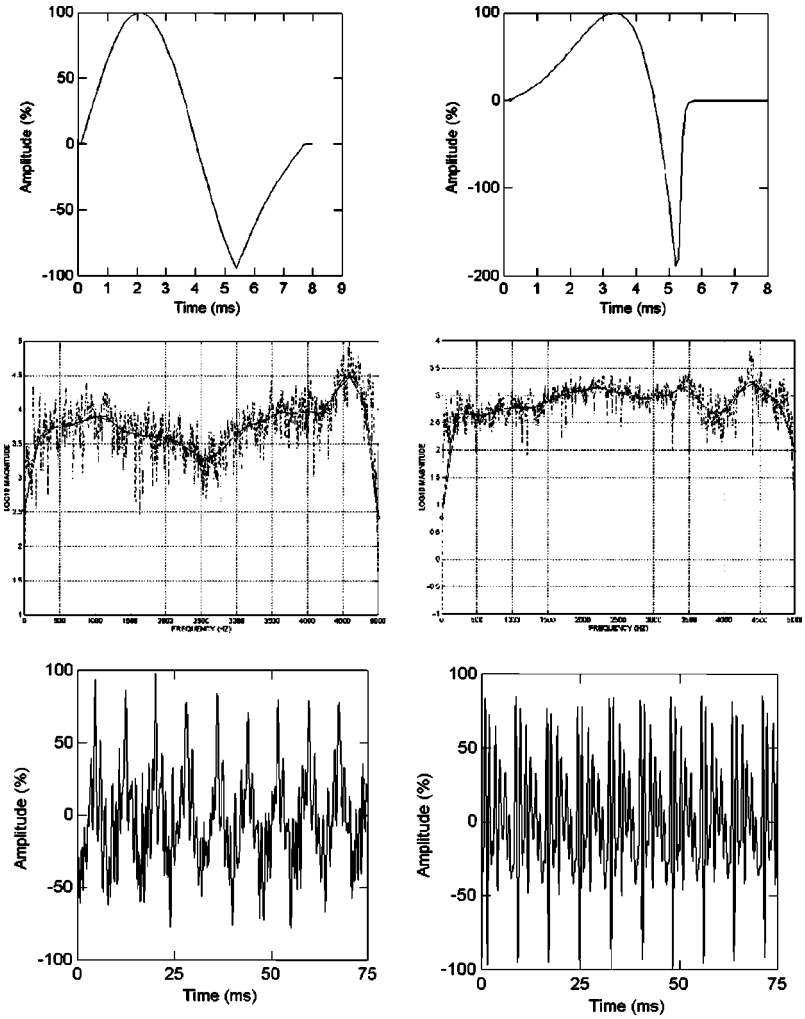

FIG. 4. Sources, noise spectra, and time series waveforms for representative stimuli used in experiment 2. The leftmost three panels represent a voice with a relatively sinusoidal source and a high noise-to-signal ratio (−6.9 dB). The rightmost three panels represent a voice with a nonsinusoidal source and a lower noise-to-signal ratio (−40.4 dB).

$H1$-$H2$, the more sinusoidal the source. Variability in NSR responses increased as $H1$-$H2$ decreased ($r=-0.57$, $p<0.02$). Listeners agreed best in their NSR responses when the harmonic part of the voicing source was near-sinusoidal, as predicted by theory and previous research (Gerratt and Kreiman, 2001). That is, listeners were less sensitive to changes in the NSR when enough energy from the harmonic part of the voice source was also present to provide perceptually salient excitation in the high-frequency part of the voice spectrum. No such relationship was observed between harmonic source characteristics and agreement levels for jitter or shimmer (jitter: $r=0.29$, n.s.; shimmer: $r=0.28$, n.s.).

D. Discussion

These data suggest that listeners are relatively sensitive to the overall extent of aperiodicity in the signal, as measured by the NSR. Response variability was relatively low for NSR responses; responses were consistent across experimental conditions, and agreement increased with the amount of noise present. In contrast, the correspondence between jitter, shimmer, and perceived vocal quality appears far less precise. Responses varied widely when listeners were asked to match the amounts of jitter and shimmer present in natural voice signals. Jitter and shimmer responses varied significantly with experimental condition, and variability in responses increased with increasing aperiodicity. No effect of task on variability was observed, suggesting that listener difficulties were not related to the presence or absence of noise (and corresponding differences in the naturalness of the stimuli).

The dependence of jitter and shimmer responses on experimental condition suggests that listeners judge aperiodicity with respect to the overall amount of noise in the signal, rather than by decomposing aperiodicity into independent, separable aspects. This in turn suggests that listeners' relative insensitivity to the amounts of jitter and shimmer present in a voice signal may be due to their inability to isolate jitter and shimmer within a noisy voice signal. The increasing variability in jitter and shimmer responses as the NSR increased is consistent with this explanation. When the NSR was low, listeners chose (relatively) low values for all aspects of aperiodicity. As the NSR increased, however, variability in jitter and shimmer responses increased, suggesting that listeners as a group could not decide how much of the aperiodicity they heard was jitter or shimmer.

The following experiment tested this hypothesis. We synthesized stimuli with and without noise, and added jitter and shimmer in steps to those stimuli to create several series of voices. If listeners are able to separate the effects of jitter and shimmer from the overall noise pattern, then they should be equally able to detect differences between these stimuli

TABLE I. Stimulus characteristics, experiment 2.

Voice	Speaker sex	Source	NSR (dB)
1	F	−sine	−23.2
2	M	+sine	−6.9
3	F	+sine	−14.9
4	M	−sine	−40.4
5	M	+sine	−24.9
6	F	−sine	−29.2
7	F	+sine	−28.5
8	M	−sine	−29.5

whether noise is absent or present. If listeners are simply grossly insensitive to jitter and shimmer, then they should have difficulty detecting differences whether noise is present or absent. Finally, if listeners can easily distinguish among the variable jitter and shimmer responses from experiment 1, this would implicate difficulties with the multivariate matching task as the cause of the observed variability in responses.

III. EXPERIMENT 2

A. Method

1. Stimuli

Eight voices (four males and four females) were selected from the set of 20 studied in experiment 1. Because sensitivity to the level of noise in a voice may depend on the spectral shape of the harmonic part of the source (Gerratt and Kreiman, 2001), voices were chosen based on the shape of the LF-fitted source pulses (relatively sinusoidal/relatively nonsinusoidal) and on the NSR of the natural voice (relatively high/relatively low) (Fig. 4). Stimuli were chosen so that these parameters varied as orthogonally as possible, given the characteristics of the original voice samples. One male and one female voice were assigned to each source-by-NSR cell (Table I).

Five series of stimuli, each comprising five tokens, were synthesized for each of these eight voices. In two series, the amount of jitter increased across the five tokens in steps. One jitter series was synthesized with the NSR set at a constant value equal to the mean of the levels listeners selected for that voice in experiment 1. The second jitter series was synthesized with the NSR set at −50 dB (no perceptible noise present). Two series of shimmered stimuli were also created, one in which the NSR was set at −50 dB and one with the NSR set at the mean of the values listeners selected in experiment 1. In both series, the amounts of shimmer present increased across the series in steps, as for the jittered stimuli. The fifth series of stimuli was synthesized with the amount of noise increasing in steps, without any additional jitter or shimmer. Because noise responses were independent of jitter and shimmer responses in experiment 1, additional stimulus series were not created varying noise levels in the contexts of average amounts of jitter and shimmer.

Series endpoints and step sizes were individually determined for each voice (Table II). Endpoints for the jitter series represented the maximum and minimum jitter responses observed for that voice in the condition in experiment 1 in which listeners adjusted jitter only, with noise and shimmer set at −50 and 0 dB, respectively. The three intermediate points in each series were evenly spaced (in acoustic units) between these extremes. Endpoints for the shimmer and noise series were similarly selected based on ''shimmer only'' and ''noise only'' response ranges from experiment 1. All other synthesis parameters were held constant for each voice at the values used in experiment 1.

Stimuli were 1 s in duration, and were synthesized with a sampling rate of 10 kHz, using the methods described for experiment 1. They were multiplied by 25-ms onset and offset ramps and scaled for equal peak amplitude prior to presentation to listeners.

2. Listeners and listening task

Eighteen listeners participated in this experiment. All reported normal hearing. For each series of stimuli for each voice, listeners heard all possible pairs of the five synthetic tokens in the series, plus an equal number of pairs where stimuli were the same, for a total of 800 trials/listener (8 voices×5 series/voice×10 comparisons/series, plus 400 ''voices same'' trials). They were asked to determine whether the stimuli were the same or different within a pair, and to rate their confidence in their response on a 5-point scale ranging from ''positive'' to ''wild guess.''

Testing took place in a double-walled sound suite in two sessions, each lasting about 45 min. Stimuli were presented in free field at a comfortable constant listening level, and were rerandomized for each listener. Voices within a pair were separated by 350 ms. Listeners controlled the rate at which pairs were presented, but were not allowed to play a pair more than once before responding.

TABLE II. Characteristics of stimulus series for jitter, shimmer, and noise in experiment 2.

Voice	Step sizes for jitter (%)	Jitter continuum endpoints	Step sizes for shimmer (dB)	Shimmer continuum endpoints	Step sizes for NSR (dB)	NSR continuum endpoints
1	0.26	0.018, 1.05	0.18	0, 0.73	1.60	−27.15, −20.80
2	0.65	0.15, 2.75	0.39	0.18, 1.75	1.46	−10.10, −4.25
3	0.41	0, 1.63	0.45	0, 1.82	0.87	−16.90, −13.40
4	0.25	0, 0.98	0.10	0, 0.41	4.68	−49.20, −30.45
5	0.47	0.08, 1.94	0.30	0, 1.18	1.97	−29.20, −21.30
6	0.19	0, 0.74	0.15	0.04, 0.64	4.57	−41.95, −23.65
7	0.23	0, 0.90	0.24	0.09, 1.07	4.17	−40.20, −23.50
8	0.70	0, 2.80	0.29	0, 1.16	2.84	−36.60, −25.25

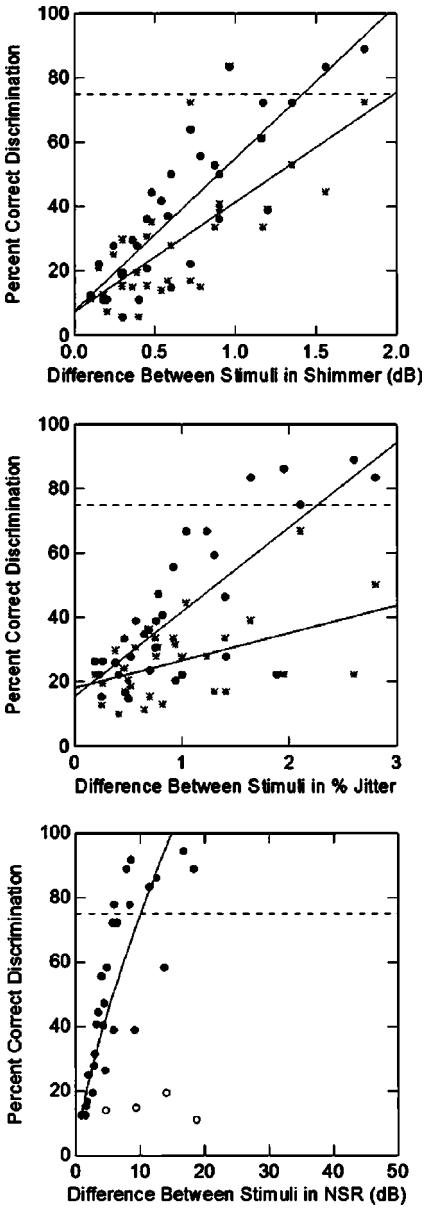

FIG. 5. The rate of correct ''different'' responses versus as a function of the acoustic distance between stimuli. For jitter and shimmer data, stimuli from series that included noise are plotted with stars, and stimuli from series synthesized without noise are plotted with filled circles. For noise data, open circles indicate a voice with a nonsinusoidal source and a high NSR. Regression lines through the two series of data in each panel are plotted with solid lines. The dashed line corresponds to 75% correct discrimination, as described in the text. Top panel: shimmer. Middle panel: jitter. Bottom panel: the noise-to-signal ratio.

3. Results and discussion

Same/different response data were pooled across listeners to estimate overall discrimination performance.[3] For each voice and stimulus series, we calculated the percentage of correct ''same'' responses and the percentage of correct ''different'' responses. The rate of correct ''same'' responses was consistent across experimental conditions [$F(4,35) = 0.14$, $p > 0.05$], ranging from 78.9%–92.8% (mean =86.5%, sd=3.70%). Discrimination accuracy (the rate of correct ''different'' responses) is shown as a function of the acoustic distance between stimuli within a pair in Fig. 5. Each panel of this figure combines data from the eight individual stimulus voices, pooled across listeners.

Results for jitter and shimmer are shown in panels (A) and (B). In these panels, stimuli from series that included noise are plotted with stars, and stimuli from series synthesized without noise are plotted with filled circles. Regression lines through the two series of data in each panel are plotted with solid lines [Jitter: with noise, $F(1,30) = 9.12$, $p < 0.01$; no noise, $F(1,30) = 46.39$, $p < 0.01$. Shimmer: with noise, $F(1,30) = 32.33$, $p < 0.01$; no noise, $F(1,30) = 83.56$, $p < 0.01$]. Data in panels (A) and (B) show similar patterns. Discrimination accuracy was significantly better in both cases when noise was absent (filled circles) than when it was present (stars). For jitter, correct discrimination rates averaged 14.95% greater when noise was absent than when it was present [matched pairs $t(31) = 4.37$, $p < 0.01$]. For shimmer, discrimination accuracy averaged 9.03% more when noise was absent [matched pairs $t(31) = 3.74$, $p < 0.01$].

Difference limens were estimated by the point of intersection between the regression lines fit to the data and the point at which discrimination accuracy reached 75% (indicated by a dashed line in the figures; e.g., Marks and Algom, 1998). For jitter, averaged across the eight voices, listeners only reliably heard a difference between stimuli that differed by 2.27% or more in jitter when spectral noise was absent. When noise was present, listeners never reliably heard a difference between the stimuli used in this experiment, indicating that stimuli must differ by more than 3% jitter (the maximum range available) for the difference to be reliably perceptible. For shimmer, in the absence of spectral noise listeners could discriminate only among stimuli that differed by at least 1.42 dB in mean shimmer; in the context of spectral noise, a difference of 1.99 dB in mean shimmer (given a maximum range of 2.0 dB) was required for listeners to hear a difference between stimuli 75% of the time.[4]

The pattern is somewhat different for spectral noise [panel (C)]. One voice (number 4) with an extremely high NSR and a harmonic source that provided significant high-frequency excitation was an obvious outlier (plotted with open circles in the figure). When this voice is omitted from the analysis, the function relating discrimination accuracy to differences in noise levels appears curvilinear. This curve was interpolated using a power function rather than a linear function [$F(1,26) = 43.58$, $p < 0.01$], as shown in the figure, omitting the outlying case. For stimuli to be correctly discriminated 75% of the time required a difference in the NSR of 10.65 dB, a relatively low value given the 50-dB range of the scale. (Recall that listeners used the entire scale for all three measures, so observed differences in the magnitude of the estimated difference limens with respect to the length of the scales are not due to the length of the scales themselves.)

Experiment 2 examined three hypotheses regarding the causes of poor listener agreement in experiment 1. The present results are consistent with two of these hypotheses: Listeners are unable to isolate jitter and shimmer as separate components in the overall pattern of aperiodicity in a voice, and they are also insensitive overall to jitter and shimmer. Listeners were significantly better at discriminating levels of both jitter and shimmer when noise was absent than when it

was present, indicating that listeners do have difficulty decomposing the overall aperiodic component of a voice into independent constituent parts. However, estimated difference limens for jitter and shimmer are large overall, both in the presence and in the absence of spectral noise. These limens are also large relative to jitter and shimmer levels usually considered experimentally or clinically significant. For example, Karnell *et al.* (1995) reported that differences in the jitter measured by different analysis systems averaged 0.01%, while differences in shimmer averaged 0.085 dB. Bielamowicz *et al.* (1996) reported differences between systems in measured jitter of about 0.4%–0.5%; differences between systems in measured shimmer values were less than 0.1 dB. Linville (2000) reported that the voices of old and young men differed in jitter by about 0.6%. Shimmer values for old and young men differed by about 0.4 dB. Hanson *et al.* (1997) found decreases in jitter of about 0.03% after 6–9 weeks of treatment for laryngeal inflammation from gastroesophageal reflux, and claimed that these changes ''document the changes in vocal quality with treatment for chronic laryngitis'' (p. 284). Shimmer levels for patients with unilateral vocal-fold paralyses differed from those of control patients by about 1 dB (Hartl *et al.*, 2003); and reliable changes in jitter level during a histamine challenge test averaged about 1.5%, leading to the conclusion that ''jitter is an objective and repeatable measurement of hoarseness'' (Jones *et al.*, 2001, p. 29). Although probable variations in measurement techniques may limit comparisons among studies, estimated difference limens for jitter and shimmer in the present study far surpass these values, and also exceed the measurement precision usually required for such measures (Titze, 1995; Titze *et al.*, 1987).

Finally, the results do not support the hypothesis that listener problems in agreement observed in experiment 1 resulted from difficulties in performing the multivariate method-of-adjustment task. In our previous study using this method (Gerratt and Kreiman, 2001), listener agreement was substantially higher for NSR settings than it was for perceptual ratings of the noisiness of the same stimuli. However, in that study listeners were asked to adjust only a single synthesizer parameter. Similar levels of listener agreement were observed for NSR settings in the present study, when listeners were asked to adjust as many as three parameters simultaneously. Further, NSR ratings remained consistent, whether listeners were adjusting one, two, or three parameters simultaneously. The fact that listener agreement about NSR ratings remained consistently high across tasks suggests that listeners' inconsistency with each other when matching jitter and shimmer levels was not due to the multivariate task, but instead was related to their auditory insensitivity to these two parameters.

IV. GENERAL DISCUSSION

Acoustic measures of voice derive their importance from the relevance of the acoustic signal to auditory perception of voice, association with some aspect of laryngeal physiology, or both (Catford, 1977). Jitter, shimmer, and noise have many physical sources and may arise at many stages in the speech production process (see, e.g., Titze, 1994, for review). Because a given measurement value can reflect so many different causes, acoustic measures of aperiodicity are not diagnostically useful as indices of any particular physical state or physiological process. This leaves associations with voice quality as a possible motivation for measuring aperiodicity. This motivation also appears to fail in the cases of jitter and shimmer. Comparisons with published values are difficult because recording techniques, measurement procedures, and computational algorithms vary widely (see Buder, 2000, for review), and authors are not always specific about the methods applied. However, it appears that listeners are insensitive to the amounts of jitter and shimmer present in a voice sample within a range often treated as meaningful, as discussed above. We conclude that the associations between jitter, shimmer, and perceived voice quality are not sufficiently explanatory to justify continued reliance on jitter and shimmer as indices of voice quality.

The case is stronger for the perceptual relevance of NSR measures, because listeners agreed far better in their NSR responses than they did for jitter and shimmer, and the minimum reliably perceptible difference for the NSR was much smaller relative to the range of observations than it was for jitter and shimmer. NSR responses were independent of experimental task, and observed variability could be explained in part by the pattern of high-frequency excitation in a voice. These results indicate that listeners respond perceptually to changes in the NSR in consistent and principled ways, suggesting that the NSR is a significant and reliable determinant of vocal quality.

However, the fact that a measure is perceptually or psychologically important does not mean that researchers can ignore the limits of perceptual resolution on that scale when applying it. Although comparisons to values in the literature are again difficult (differences among computational methods for NSR measures being particularly inscrutable and vexing; see Buder, 2000), many authors' claims about group differences or treatment effects rest on NSR differences that may be imperceptible. For example, Su *et al.* (2002) found statistically significant differences in pre- and postsurgical NSR values of 0.12 dB; Jotz *et al.* (2002) reported that each increase of 0.01 dB in the NSR doubled the risk of dysphonia in a sample of boys with and without vocal-fold lesions; and Niedzielska (2001) found differences in NSRs between control subjects and various diagnostic groups ranging from 2.7 to 14.2 dB. Further research examining the perceptual interactions between the harmonic and inharmonic parts of the voicing source should contribute to standardizing NSR measures so that they reflect vocal quality as accurately as possible. This will enhance our ability to apply NSR measures appropriately.

Finally, the present data extend our previous findings regarding the reliability of the method of adjustment task from expert listeners (Gerratt and Kreiman, 2001) to naive listeners. Reanalysis of data from our previous experiment, which used a different set of stimulus voices and expert listeners, produced a difference limen for the NSR (the point at which listeners achieved 75% correct discrimination, on average) of 13.4 dB, compared to the value of 10.65 dB for naive listeners in the present study. Expert and naive listeners

have previously been shown to differ significantly in the perceptual strategies they apply when rating pathological voices on traditional scales like breathiness or roughness (Kreiman *et al.*, 1990). However, the similarity in sensitivity levels for expert and naive listeners suggests that this method-of-adjustment task controls the effects of listener experience on perceptual responses, as previously predicted (Kreiman and Gerratt, 2000).

The results reported here may be surprising in light of the hundreds of papers published on these acoustic measures over the last 40 years. These results highlight the importance of applying psychometric methods to the study of voice quality. Descriptive and correlational statistical techniques, including multidimensional scaling and factor analysis, may suggest that a given dimension is perceptually important. However, without confirmatory experimental studies, such associations between signal and percept remain merely suggestive. At a minimum, both theory and application will benefit if investigators verify that listeners are auditorily sensitive to the range of values of interest for a particular acoustic measure to ensure that such measures are in fact determinants of perceived quality.

ACKNOWLEDGMENTS

Norma Antoñanzas and Brian Gabelman wrote the software used in this research. Jason Mallory created the stimuli in experiment 2 and patiently tested many of the listeners in both experiments. Preliminary reports of the results were presented at the 144th and 145th meetings of the Acoustical Society of America. This research was supported by Grant DC01797 from the National Institute on Deafness and Other Communication Disorders.

[1]Amplitude modulations did not produce major differences in the quality of most voices, because many so-called ''amplitude tremors'' are artifacts of frequency modulation (Sundberg, 1995; Kreiman *et al.*, 2003). However, these modulations were important for successfully synthesizing several samples in the present study, especially those from speakers with spasmodic dysphonia.

[2]Software is available from the authors by request.

[3]Confidence ratings were not used in analyses of these data.

[4]Difference limens observed here may depend in part on the fact that listeners heard stimuli in free field, and could only hear a pair of voices once before responding. Although results reflect listeners' discrimination ability under normal listening conditions, they do not necessarily reflect the precise limits of perceptual acuity as measured with headphones and multiple presentations prior to response.

Andrianopoulos, M. V., Darrow, K. N., and Chen, J. (**2001**). ''Multimodal standardization of voice among four multicultural populations: Fundamental frequency and spectral characteristics,'' J. Voice **15**, 194–219.

ANSI (**1960**). ANSI S1.1-1960, ''Acoustical terminology'' (American National Standards Institute, New York).

Bielamowicz, S., Kreiman, J., Gerratt, B. R., Dauer, M. S., and Berke, G. S. (**1996**). ''A comparison of voice analysis systems for perturbation measurement,'' J. Speech Hear. Res. **39**, 126–134.

Buder, E. H. (**2000**). ''Acoustic analysis of voice quality: A tabulation of algorithms 1902–1990,'' in *Voice Quality Measurement*, edited by R. D. Kent and M. J. Ball (Singular, San Diego), pp. 119–244.

Catford, J. C. (**1977**). *Fundamental Problems in Phonetics* (Indiana University Press, Bloomington, IN).

Coleman, R. F., and Wendahl, R. W. (**1967**). ''Vocal roughness and stimulus duration,'' Speech Monographs **34**, 85–92.

Deal, R., and Emanuel, F. W. (**1978**). ''Some waveform and spectral features of vowel roughness,'' J. Speech Hear. Res. **21**, 250–264.

de Krom, G. (**1993**). ''A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals,'' J. Speech Hear. Res. **36**, 254–266.

de Krom, G. (**1995**). ''Some spectral correlates of pathological breathy and rough voice quality for different types of vowel fragments,'' J. Speech Hear. Res. **38**, 794–811.

Fant, G., Liljencrants, J., and Lin, Q. (**1985**). ''A four-parameter model of glottal flow,'' STL-QPSR **4**, 1–13.

Fukazawa, T., El-Assuooty, A., and Honjo, I. (**1988**). ''A new index for evaluation of the turbulent noise in pathological voice,'' J. Acoust. Soc. Am. **83**, 1189–1193.

Gabelman, B. (**2003**). ''Analysis and synthesis of pathological vowels,'' unpublished doctoral dissertation, Department of Electrical Engineering, University of California at Los Angeles.

Gerratt, B. R., and Kreiman, J. (**2001**). ''Measuring voice quality with speech synthesis,'' J. Acoust. Soc. Am. **110**, 2560–2566.

Green, D. M., and Swets, J. A. (**1966**). *Signal Detection Theory and Psychophysics* (Krieger, Huntington, NY).

Hanson, D. G., Jiang, J. J., Chen, J., and Pauloski, B. R. (**1997**). ''Acoustic measurement of change in voice quality with treatment for chronic posterior laryngitis,'' Ann. Otol. Rhinol. Laryngol. **106**, 279–285.

Hartl, D. A., Hans, S., Vaissiere, J., and Brasnu, D. A. (**2003**). ''Objective acoustic and aerodynamic measures of breathiness in paralytic dysphonia,'' Eur. Arch. Otorhinolaryngol. **260**, 175–182.

Heiberger, V. L., and Horii, Y. (**1982**). ''Jitter and shimmer in sustained phonation,'' in *Speech and Language: Advances in Basic Research and Practice*, edited by N. J. Lass (Academic, New York), Vol. 7, pp. 299–332.

Hillenbrand, J. (**1987**). ''A methodological study of perturbation and additive noise in synthetically generated voice signals,'' J. Speech Hear. Res. **30**, 448–461.

Hillenbrand, J. (**1988**). ''Perception of aperiodicities in synthetically generated voices,'' J. Acoust. Soc. Am. **83**, 2361–2371.

Javkin, H., Antonanzas-Barroso, N., and Maddieson, I. (**1987**). ''Digital inverse filtering for linguistic research,'' J. Speech Hear. Res. **30**, 122–129.

Jensen, P. J. (**1965**). ''Adequacy of terminology for clinical judgment of voice quality deviation,'' Eye Ear Nose Throat Mon. **44**, 77–82.

Jones, T. M., Trabold, M., Plante, F., Cheetham, B. M., and Earis, J. E. (**2001**). ''Objective assessment of hoarseness by measuring jitter,'' Clin. Otolaryngol. **26**, 29–32.

Jotz, G. P., Cervantes, O., Abrahao, M., Settanni, F. A., and de Angelis, E. C. (**2002**). ''Noise-to-harmonics ratio as an acoustic measure of voice disorders in boys,'' J. Voice **16**, 28–31.

Karnell, M. P., Hall, K. D., and Landahl, K. L. (**1995**). ''Comparison of fundamental frequency and perturbation measurements among three analysis systems,'' J. Voice **9**, 383–393.

Kreiman, J., and Gerratt, B. R. (**1996**). ''The perceptual structure of pathologic voice quality,'' J. Acoust. Soc. Am. **100**, 1787–1795.

Kreiman, J., and Gerratt, B. R. (**2000**). ''Measuring vocal quality,'' in *Voice Quality Measurement*, edited by R. D. Kent and M. J. Ball (Singular, San Diego), pp. 73–102.

Kreiman, J., Gabelman, B., and Gerratt, B. R. (**2003**). ''Perception of vocal tremor,'' J. Speech Lang. Hear. Res. **46**, 203–214.

Kreiman, J., Gerratt, B. R., and Berke, G. S. (**1994**). ''The multidimensional nature of pathologic vocal quality,'' J. Acoust. Soc. Am. **96**, 1291–1302.

Kreiman, J., Gerratt, B. R., and Precoda, K. (**1990**). ''Listener experience and perception of voice quality,'' J. Speech Hear. Res. **33**, 103–115.

Kreiman, J., Vanlancker-Sidtis, D., and Gerratt, B. R. (**2005**). ''Perception of voice quality,'' to appear in *Handbook of Speech Perception*, edited by D. B. Pisoni and R. E. Remez (Blackwell, Oxford), pp. 338–362.

Linville, S. E. (**2000**). ''The aging voice,'' in *Voice Quality Measurement*, edited by R. D. Kent and M. J. Ball (Singular, San Diego), pp. 359–376.

MacMillan, N. A., and Creelman, C. D. (**1991**). *Detection Theory: A User's Guide* (Cambridge University Press, Cambridge).

Marks, L. E., and Algom, D. (**1998**). ''Psychological scaling,'' in *Measurement, Judgment, and Decision Making*, edited by M. H. Birnbaum (San Diego, Academic), pp. 81–178.

Martin, D., Fitch, J., and Wolfe, V. (**1995**). ''Pathologic voice type and the acoustic prediction of severity,'' J. Speech Hear. Res. **38**, 765–771.

Michaelis, D., Gramss, T., and Strube, H. W. (**1997**). ''Glottal-to-noise excitation ratio—A new measure for describing pathological voices,'' Acustica **83**, 700–706.

Munoz, J., Mendoza, E., Fresneda, M. D., Carballo, G., and Lopez, P. (**2003**). ''Acoustic and perceptual indicators of normal and pathological voice,'' Folia Phoniatr Logop **55**, 102–114.
Murphy, P. J. (**1999**). ''Perturbation-free measurement of the harmonics-to-noise ratio in voice signals using pitch synchronous harmonic analysis,'' J. Acoust. Soc. Am. **105**, 2866–2881.
Niedzielska, G. (**2001**). ''Acoustic analysis in the diagnosis of voice disorders in children,'' Int. J. Pediatr. Otorhinolaryngol. **57**, 189–193.
Prosek, R. A., Montgomery, A. A., Walden, B. E., and Hawkins, D. B. (**1987**). ''An evaluation of residue features as correlates of voice disorders,'' J. Commun. Disord. **20**, 105–117.
Su, C. Y., Lui, C. C., Lin, H. C., Chiu, J. F., and Cheng, C. A. (**2002**). ''A new paramedian approach to arytenoid adduction and strap muscle transposition for vocal fold medialization,'' Laryngoscope **112**, 342–350.
Sundberg, J. (**1995**). ''Acoustic and psychoacoustic aspects of vocal vibrato,'' in *Vibrato*, edited by P. H. Dejonckere, M. Hirano, and J. Sundberg (Singular, San Diego, CA), pp. 35–62.
Titze, I. R. (**1994**). *Principles of Voice Production* (Prentice Hall, Englewood Cliffs, NJ).
Titze, I. R. (**1995**). *Workshop on Acoustic Voice Analysis Summary Statement* (National Center for Voice and Speech, Denver).
Titze, I. R., Horii, Y., and Scherer, R. C. (**1987**). ''Some technical considerations in voice perturbation measurements,'' J. Speech Hear. Res. **30**, 252–260.
Wendahl, R. W. (**1963**). ''Laryngeal analog synthesis of harsh voice quality,'' Folia Phoniatr. **15**, 241.
Wendahl, R. W. (**1966a**). ''Some parameters of auditory roughness,'' Folia Phoniatr. **18**, 26–32.
Wendahl, R. W. (**1966b**). ''Laryngeal analog synthesis of jitter and shimmer auditory parameters of harshness,'' Folia Phoniatr. **18**, 98–108.