Measuring vocal quality with speech synthesis

Bruce R. Gerratt[a] and Jody Kreiman[b]

Division of Head and Neck Surgery, UCLA School of Medicine, 31-24 Rehab Center, Los Angeles, California 90095-1794

(Received 7 December 2000; accepted for publication 13 August 2001)

Much previous research has demonstrated that listeners do not agree well when using traditional rating scales to measure pathological voice quality. Although these findings may indicate that listeners are inherently unable to agree in their perception of such complex auditory stimuli, another explanation implicates the particular measurement method—rating scale judgments—as the culprit. An alternative method of assessing quality—listener-mediated analysis-synthesis—was devised to assess this possibility. In this new approach, listeners explicitly compare synthetic and natural voice samples, and adjust speech synthesizer parameters to create auditory matches to voice stimuli. This method is designed to replace unstable internal standards for qualities like breathiness and roughness with externally presented stimuli, to overcome major hypothetical sources of disagreement in rating scale judgments. In a preliminary test of the reliability of this method, listeners were asked to adjust the signal-to-noise ratio for 12 synthetic pathological voices so that the resulting stimuli matched the natural target voices as well as possible For comparison to the synthesis judgments, listeners also judged the noisiness of the natural stimuli in a separate task using a traditional visual-analog rating scale. For 9 of the 12 voices, agreement among listeners was significantly (and substantially) greater for the synthesis task than for the rating scale task. Response variances for the two tasks did not differ for the remaining three voices. However, a second experiment showed that the synthesis settings that listeners selected for these three voices were within a difference limen, and therefore observed differences were perceptually insignificant. These results indicate that listeners can in fact agree in their perceptual assessments of voice quality, and that analysis-synthesis can measure perception reliably. *© 2001 Acoustical Society of America.*

[DOI: 10.1121/1.1409969]

PACS numbers: 43.71.Bp, 43.71.Gv [KRK]

I. INTRODUCTION

The study of voice quality has always been challenging. Difficulties arise at the definitional level (e.g., Jensen, 1965; Sundberg, 1987) and compound at every step of investigation. The links between perceived quality and the underlying vocal physiology or acoustics are not well understood, so the importance of perceptual voice features cannot be established by reference to objective measures of phonatory function or the acoustic waveform. The inherently multidimensional nature of voice quality increases the complexity of measurement and interpretation of results. Further, no satisfactory basis exists for defining an *a priori* structure for the perceptual space, and the nature of a set of features that might specify voice quality remains unknown. Because of this, instrumental measures of voice cannot be validated in a straightforward manner by their relationship to independently validated perceptual constructs. Finally, the appropriate method for measuring what listeners hear remains an unresolved issue. Research on voice quality perception has used rating scale techniques (including visual analog scales, equal-appearing interval scales, and direct magnitude estimation) almost exclusively. However, listeners often disagree in their assessment of overall voice quality and in their use of scales for individual voice qualities like roughness and breathiness (Gelfer, 1988; Kreiman and Gerratt, 1998; Orlikoff, 1999; Wuyts *et al.*, 1999; but cf. Hillenbrand *et al.*, 1994; Hillenbrand and Houde, 1996). [See, e.g., Poulton (1979) and Gescheider (1997) for review of similar issues regarding judgments of the loudness of pure tones.] Difficulty isolating single perceptual dimensions of complex, multidimensional stimuli appears to be a significant source of listener disagreement (Kreiman and Gerratt, 2000; Kreiman *et al.*, 1994). Such listener disagreement reduces confidence in the validity and utility of existing perceptual measures of voice, further undermining attempts to define, describe, or quantify vocal quality.

Finding valid and reliable alternatives to traditional voice quality scaling methods requires hypotheses about the sources of listener disagreements, so that techniques can be developed to control such variability. Previous studies of pathological voices (Gerratt *et al.*, 1993; Kreiman and Gerratt, 2000) suggest that traditional perceptual scaling methods are effectively matching tasks, where external stimuli (the voices) are compared to stored mental representations that serve as internal standards for the various rating scales. These idiosyncratic, internal standards appear to vary with listeners' previous experience with voices (Kreiman *et al.*, 1990; Verdonck-de Leeuw, 1998) and with the context in which a judgment is made (Gerratt *et al.*, 1993; cf. Gescheider and Hughson, 1991), and may vary substantially across listeners as well as within a given listener (Gerratt

[a] Electronic mail: bgerratt@ucla.edu

[b] Electronic mail: jkreiman@ucla.edu

et al., 1993; Kreiman *et al.*, 1993). In addition, severity of vocal deviation, difficulty isolating individual dimensions in complex perceptual contexts, and factors like lapses in attention can also influence perceptual measures of voice (de Krom, 1994; Kreiman and Gerratt, 2000). These factors (and possibly others) presumably all add uncontrolled variability to scalar ratings of vocal quality, and contribute to listener disagreement.

This article describes a preliminary test of an alternative approach to the assessment of voice quality that is designed to control these sources of measurement error. In this method, listeners vary speech synthesis parameters to create an acceptable auditory match to a natural voice stimulus. When a listener chooses a best match to a test stimulus, the synthesis settings parametrically represent the listener's perception of voice quality. Because listeners directly compare each synthetic token they create to the target natural voice, they need not refer to internal standards for particular voice qualities. Further, listeners can manipulate acoustic parameters and hear the result of their manipulations immediately. We hypothesize that this process will help listeners focus attention on individual acoustic dimensions, reducing the perceptual complexity of the assessment task and presumably the associated response variability. In theory, then, the analysis-synthesis method should improve agreement among listeners in their assessments of voice quality relative to traditional rating scale techniques, because it controls the major sources of variance in quality judgments.

Similar logic motivated a previous attempt to reduce variability in ratings of voice quality through the use of fixed external reference stimuli (perceptual ''anchors'') (Gerratt *et al.*, 1993). Listeners in that study assessed vocal roughness of synthetic stimuli using a five-point scale in which a synthetic voice sample exemplified each scale value. Because listeners compared test stimuli to these ''anchor'' stimuli, we hypothesized that they would agree better in their ratings than they would when referring only to their internal criteria for different levels of roughness in a traditional rating scale task. In fact, listener agreement did increase significantly when test stimuli were identical to or immediately adjacent to the anchor stimuli, but agreement decreased sharply for stimuli that fell further from the anchors in the series [cf. Wedell *et al.* (1990), who found similar effects in ratings of the severity of psychiatric symptoms]. Presumably, the further apart the test and anchor stimuli were acoustically, the more listeners relied on their internal quality standards, resulting in lower interrater agreement. Thus, while the anchored task reduced overall variation in listener ratings of voice quality, the limited scale resolution proved a significant source of measurement error.

The analysis-synthesis technique described in this report is designed to correct this limitation. This task provides the same theoretical advantages as the anchored protocol, in that listeners explicitly match reference and test stimuli. However, the analysis-synthesis task provides much finer scale resolution, allowing listeners to create a very close match to the perceived quality of the test voice. Anchored protocols also require the experimenter to choose the specific unidimensional quality scales along which voices are to be rated (e.g., breathiness, roughness, etc.). Such scales have resisted empirical validation, and the nature of the ''correct'' set of scalar features for voice quality remains unknown. In contrast, analysis-synthesis techniques permit assessment of quality without the need to postulate discrete perceptual scales for particular aspects of quality, because synthesis parameters combine to model overall vocal quality. These theoretical advantages should provide better measurement reliability and validity for a synthesis rating task relative to traditional scalar voice rating techniques. The experiment reported below provides a preliminary test of this hypothesis.

II. EXPERIMENT 1

In this preliminary test of the analysis-synthesis method of evaluating voice quality, listeners were asked to manipulate a single acoustic parameter. Although voice quality is multidimensional, restricting the task in this way simplifies comparing the reliability of the synthesis task to that of traditional rating scale measures of voice, and presents the simplest case to test the potential efficacy of this technique. The signal-to-noise ratio (SNR) was selected because it has been historically important in describing voice quality (e.g., Hirano, 1988; Michaelis *et al.*, 1998; Yumoto *et al.*, 1982), and because it can be manipulated with a single parameter (unlike the periodic component of the source, for example, which must be specified with several parameters as described below). In a separate experiment, listeners also evaluated the ''noisiness'' of the stimulus voices using a traditional visual analog rating scale. To the extent that the synthesis protocol controls sources of interrater variability, listeners should agree better in their choice of synthesis settings than they do in the analogous rating scale task.

A. Method

1. Voice samples

Twelve samples of the vowel /a/ were selected from a set recorded from patients in a clinical setting. Pathological voice samples were studied for two reasons. First, accurate and reliable measurement of voice quality is of particular clinical importance, because patients and clinicians are often concerned about the nature and extent of vocal deviation. Further, pathological voices encompass a wide range of the human larynx's phonatory potential, and samples with pathology have a greater range of vocal quality variables than do samples of normal phonation. Speakers ranged in age from 26 to 73 years (mean = 46.3 years), and represented a variety of primary diagnoses, including essential vocal tremor, vocal fold mass lesions, vocal fold paralysis, adductory spasmodic dysphonia, reflux laryngitis, glottal incompetence, and laryngeal web. They ranged from mildly to severely dysphonic. Both male ($n=4$) and female speakers ($n=8$) were included.

Voice signals were transduced with a 1-in. Bruel and 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

A synthetic copy of each voice was created by an experienced operator using a custom formant synthesizer implemented in MATLAB (MathWorks, Natick, MA).[1] The synthesizer operates in near-real time, and allows users to manipulate the fundamental frequency ($f0$), the shape of the glottal volume velocity derivative, the spectrum of the inharmonic component of the voice source (the noise spectrum), the overall signal-to-noise ratio, formant frequencies and bandwidths, and the rate, extent, and regularity of frequency modulation (tremor).

The goal of synthesis at this stage of the investigation was to create the best possible copy of each target voice, so that listener performance later in the experiment could be attributed to task factors, rather than to limitations of the stimuli. Initial parameter settings for synthesis were derived from acoustic analyses of the voices as follows. Formant frequencies and bandwidths were estimated using autocorrelation linear predictive coding (LPC) analysis with a window of 25.6 ms (increased to 51.2 ms when stimulus $f0$ was near or below 100 Hz). The glottal volume velocity derivative was estimated by inverse filtering a single glottal pulse from the microphone recordings, using the method described by Javkin *et al.* (1987). The frequency of this cycle served as the initial estimate of mean $f0$.

The output of the inverse filter was least-squares fit with a 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 traditional LF model was slightly modified to improve the fit of the return phase, and so that the beginning of the closed phase was explicitly specified with a parameter. The ''equal area constraint,'' which requires that areas under the positive and negative curves in the flow derivative be the same, was also abandoned. When modeling some voices with this constraint in place, the return phases did not consistently return to zero. This introduced high-frequency artifacts when the next pulse began at 0 (see Epstein *et al.*, 1999, for further details).

Slow variations in $f0$ (vocal tremors) were modeled by modulating the nominal $f0$ in one of two patterns: a sine wave, or irregular modulation. Rate and extent of frequency modulation in both cases were estimated from $f0$ tracks of the natural voices (see Kreiman *et al.*, 2001, for details).[2]

To synthesize noise, a cepstral-domain comb filter similar to that described by de Krom (1993) first removed the harmonic part of the signal, leaving an estimate of the inharmonic component of the voice. This residual was then inverse filtered to remove the effects of vocal tract resonances, leaving the inharmonic part of the source. Next, this noise spectrum was fitted with a 25-segment piece-wise linear approximation. Finally, a 100-tap finite impulse response filter was synthesized for the fitted noise spectrum, and a spectrally shaped noise time series was created by passing white noise through this filter. The effects of jitter and shimmer were not modeled separately from overall spectral noise.

The synthesizer sampling rate was fixed at 10 kHz. To overcome quantization limits on modeling $f0$, the source time series was synthesized pulse by pulse using an interpolation algorithm. Within each pulse, samples were interpolated at exact sample instants as follows. A plot of $f0$ versus time was generated for the duration of the 1-s token to be synthesized, taking into account any modeled vocal tremor. Source pulses with frequencies dictated by the $f0$ versus time plot were calculated, and then concatenated to produce a synthetic time series. To reduce phase error, the absolute beginning and ending times of each pulse were tracked and used in the interpolation of succeeding pulses. At the beginning instant of each new pulse (which could occur at any time, including between samples), the $f0$ curve generated above was interpolated to find $f0$ for this LF pulse. Given that $f0$, the LF pulse was stretched or compressed to obtain the appropriate period, and sample points were calculated accordingly. The overall effect is equivalent to digitizing an analog pulse train with pulses of the exact desired frequencies at the fixed 10-kHz sample rate.

The LF pulse train was added to the noise time series previously described to create a complete glottal source waveform. The ratio of noise to LF energy was initially set to match the value calculated from the original voice sample. The complete synthesized source was filtered through the vocal tract model (estimated through LPC analysis, as described earlier) to generate a preliminary version of the synthetic voice. Finally, the operator adjusted all synthesis parameters to achieve the best possible perceptual match to the original voice. Note that all synthesis parameters, including noise, were optimized simultaneously. Again, the goal at this stage in the experiment was to create the best possible overall match to each target voice. In the judgment of listeners in a previous experiment (Kreiman *et al.*, 2001), all 12 synthetic stimuli provided excellent matches to the original voices. In particular, spectral noise integrated into the stimuli, resulting in very natural-sounding synthetic voices.

3. Listeners

Ten expert listeners (five otolaryngologists, three speech-language pathologists, and two phoneticians, including both authors, one of whom created the stimuli) participated in this experiment. All had extensive experience evaluating and/or treating voice disorders, and all reported normal hearing.

4. Experimental synthesis task

Prior to the experiment the signal-to-noise ratio for each synthetic stimulus was set to 50 dB. This produced synthetic voices that were free of noise. Listeners were then asked to change the overall signal-to-noise ratio until the synthetic token perceptually matched the natural target token. Listeners made their adjustments by moving a sliding cursor with a mouse along a 100-mm scale displayed on a computer monitor.[3] Left and right endpoints of this scale corresponded to signal-to-noise ratios of 50 dB (noise free) and 0 dB (high level of noise). These limits spanned the range of noise levels found in pilot studies of 70 pathological voices. Adjustments could be made in steps of 0.05 dB. Changes in the position of the cursor increased or decreased the overall

TABLE I. Rating variances for the two perceptual tasks.

Voice	Rating variance: Synthesis task	Rating variance: Visual analog rating task	*F*(9,9)
1	6.27	234.54	37.41, $p<0.05$
2	35.13	806.27	22.95, $p<0.05$
3	10.27	407.78	39.71, $p<0.05$
4	17.34	131.96	7.61, $p<0.05$
5	485.88	144.10	3.37, n.s.
6	13.97	86.46	6.19, $p<0.05$
7	33.69	157.51	4.68, $p<0.05$
8	30.97	356.27	11.50, $p<0.05$
9	433.49	317.73	1.36, n.s.
10	3.58	241.96	67.59, $p<0.05$
11	295.52	402.23	1.36, n.s.
12	11.42	554.18	48.52, $p<0.05$

signal-to-noise ratio, but did not change the shape of the noise spectrum, which remained fixed for each voice throughout the experiment. Within each trial and across all listeners, other synthesis parameters were held constant at values that produced good perceptual matches to the original voices (as determined by listeners in Kreiman *et al.*, 2001). This provided a constant perceptual frame for listeners' noise manipulations.

The experiment took place in a double-walled sound booth. Listeners heard the stimuli in free field over good quality loudspeakers, at a comfortable listening level. Voices were presented in a different random order to each listener. Listeners were able to play the synthetic token and natural target stimulus as often as necessary, and could make as many adjustments to the noise level as necessary to achieve a satisfactory match during their noise adjustments. Testing took approximately 40 min.

5. Rating scale task

Ten expert listeners (eight of whom also participated in the synthesis experiment) judged the perceived noisiness of the 12 original, natural, pathologic stimuli using a traditional 100-mm visual-analog rating scale whose two ends were labeled ''no noise'' and ''extremely noisy.'' The scale was displayed on a computer monitor, and was the same length, color, and shape as that used in the synthesis task. Listeners made their judgments by moving a sliding cursor with a mouse, as above. Listeners were able to replay the voices as often as necessary prior to making their responses. Stimuli were rerandomized for each listener. When listeners participated in both tasks, order of presentation was also randomized, and tasks were separated by at least 1 week.

B. Results

Because the signal-to-noise ratio ranged from 50 to 0, listener responses in the synthesis task were multiplied by 2 prior to statistical analysis. Thus, possible response values for both tasks ranged from 0 to 100. Table I shows the response variances for each voice, for the synthesis and visual analog rating tasks. For 9 of the 12 voices, variance in synthesis settings was significantly (and substantially) less than that for the visual analog ratings. Rating variances in the two tasks did not differ significantly for the remaining three voices.

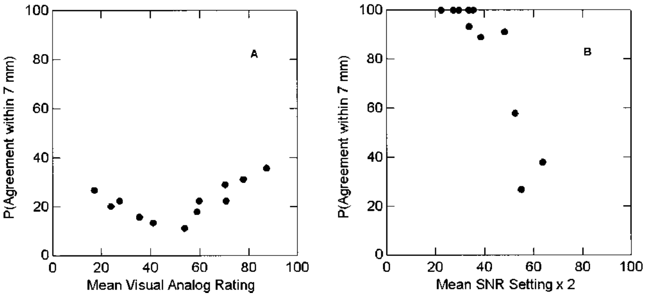

FIG. 1. For each voice, the likelihood of two raters agreeing within 7 mm in their voice assessments versus the mean rating for that voice. (a) The visual analog scale task. (b) The analysis-synthesis task. One case is hidden by overlapping points.

To compare patterns of listener agreement, we calculated the likelihood that two listeners' ratings would agree within 7 units on the 100 unit scales [corresponding to approximately 0.5 scale value on a 7 point equal-appearing interval scale (Kreiman and Gerratt, 1998)]. These probabilities are plotted against the mean rating for each voice in Fig. 1. Patterns of agreement for the visual analog task showed near zero agreement between pairs of listeners for voices whose mean rating was near the middle of the scale, with slightly better agreement for voices with mean ratings at the ends of the scale [Fig. 1(a)], similar to patterns observed for other voice quality scales (Kreiman and Gerratt, 1998). In contrast, the probability of agreement between pairs of listeners was at or near ceiling for nine stimuli in the synthesis task, and decreased for the three voices that received the highest mean SNR settings [Fig. 1(b)]. The cause of this apparent decrease in interrater reliability for these voices is examined in experiment 2, reported later in this work.

C. Discussion

These results indicate that the synthesis protocol does reduce variability in measures of perceived voice quality relative to a traditional visual analog scale task. Listeners agreed perfectly or nearly perfectly in their perceptual evaluations of pathological voice stimuli when they were given a suitable tool for reporting their judgments. In particular, listeners agreed at or near ceiling levels for the six stimuli in the midrange of mean SNR responses. To our knowledge, agreement in the midrange never occurs for traditional rating protocols (Kreiman and Gerratt, 1998). Because listener agreement improved markedly when variability due to task-related factors was controlled, listeners' disagreements in traditional rating scale tasks can be attributed to task-related difficulties, rather than to differences between listeners in perceptual processes or to the inherently subjective nature of voice perception (Weismer and Liss, 1991).

III. EXPERIMENT 2

Some of the remaining interrater variability in the synthesis task (noted particularly in 3 of the 12 voices) may be due to the lack of perceptual calibration of the signal-to-noise ratio scale. That is, differences among responses along this scale may not exceed a difference limen, and thus may not represent disagreement at all. Experiment 2 examined listeners' ability to distinguish differences in levels of noise for the stimuli in experiment 1, to determine if the apparent response variability in experiment 1 is perceptually significant.

A. Method

1. Stimuli

Twelve series of stimuli were synthesized, one series for each of the 12 voices in experiment 1, using the custom synthesizer previously described. Stimuli within each series varied only in noise levels. The minimum value for each series corresponded to the lowest level of noise selected for that voice by the listeners in experiment 1, and the maximum value corresponded to the highest level of noise selected for that voice. Four equal noise steps were created between these levels, for a total of six versions of each original voice. All other synthesis parameters were held constant within a voice "family."

Each stimulus was 1 s in duration. Stimuli were scaled for equal peak amplitude and multiplied by 30-ms ramps to eliminate click artifacts.

2. Listeners

Twenty listeners participated in this experiment. All reported normal hearing. Both expert ($n=5$) and nonexpert listeners (UCLA students; $n=15$) participated, because the task required a same/different judgment, and experts have been shown to differ from naïve listeners only for scalar ratings of individual qualities like those used in experiment 1 (e.g., Kreiman *et al.*, 1990).

3. Procedure

Listeners heard all possible pairs of the six different synthetic stimuli in each series, along with an equal number of pairs where the stimuli were identical, for a total of 360 trials/listener. Stimuli within a pair were separated by 500 ms. Stimulus pairs were rerandomized for each listener.

For each pair of voices, listeners judged whether the two stimuli were the same or different. They also rated their confidence in their response, on a five-point scale ranging from "wild guess" to "positive." These confidence ratings were used to derive receiver operating characteristics (ROCs), as described in the next section.

Testing took place in a double-walled sound booth. Stimuli were presented in free field over good quality loudspeakers. Listeners were able to replay each pair of voices as often as necessary before responding. Testing lasted approximately 45 min.

B. Results and discussion

Same/different responses were combined with listeners' confidence ratings to create a ten-point scale ranging from "certain that voices are different" to "certain that voices are the same." Receiver operating characteristics (ROCs) were constructed from these recoded data by plotting the probability of an incorrect "same" response on the x axis versus the probability of a correct "same" response on the y axis, for each level of the ten-point scale. The area under the resulting curve ranges from 0.5 (chance) to 1.0 (perfect performance) and is a measure of listeners' ability to discriminate among the different stimuli that is independent of biases in favor of "same" or "different" responses (see, e.g., Egan, 1975, or Swets and Pickett, 1982, for review).

Recall that for each six-member series of stimuli, listeners heard pairs of voices that differed in noise levels by one to five steps. To assess listeners' ability to hear larger or smaller differences in noise level, separate ROCs were constructed for stimuli differing by one step, two steps, three steps, or four to five steps, for a total of four ROCs for each of the 12 series. The 99% confidence interval around each ROC (as calculated by SPSS; SPSS Inc., Chicago, IL) was used to determine whether listeners were able to discriminate stimuli beyond chance levels. When the 99% confidence interval for the area under a given ROC included the chance value of 0.5, we concluded that differences of that size in noise steps were not perceptually significant.

No difference between expert and nonexpert listeners was found in overall discrimination accuracy [$F(1,18)=0.03$, n.s.]. To interpret the results of experiment 1, we calculated the differences between the signal-to-noise ratio levels selected by all possible pairs of raters, and evaluated the significance of each difference with respect to the ROCs derived earlier. For example, if a pair of raters differed in their chosen noise settings by between two and three noise steps, and stimuli differing by three steps were not discriminable in the present experiment, those two responses were considered to be perceptually equivalent. These analyses indicated that 72.5% of the responses produced by the listeners in experiment 1 were in fact perceptually identical (area under the ROC$\leqslant$0.5). An additional 14.2% of responses were significantly but poorly discriminable ($0.6<$area under the ROC <0.7; hit rate=86.6%; false alarm rate=58.0%, where a hit was defined as a correct "same" response). Only 3 listener-selected SNR values out of the 120 in experiment 1 (2.5%) resulted in stimuli that were consistently discriminable from the other members of their voice "families" (area under the ROC$\geqslant$0.9). These three signal-to-noise ratio responses were generated by different listeners, and were intended to match different target voices, supporting the view that they represent random rather than systematic errors (perhaps due to lapses in attention).

Recall that for three voices in experiment 1 (Nos. 5, 9, and 11; Table I), ratings from the synthesis and visual analog tasks did not differ in variance, and the likelihood of close agreement between pairs of listeners for these three voices in the synthesis task was relatively poor. ROC analyses for these voices indicated that, for voices 5 and 9, all tokens were perceptually identical (area under the ROC$\leqslant$0.5), despite the differences in signal-to-noise ratios. For voice 11, nine tokens were indiscriminable, while the tenth was only poorly discriminable from the others (area under the ROC $=0.63$). Thus, the majority of apparent listener disagreements in the analysis-synthesis task in fact resulted from the

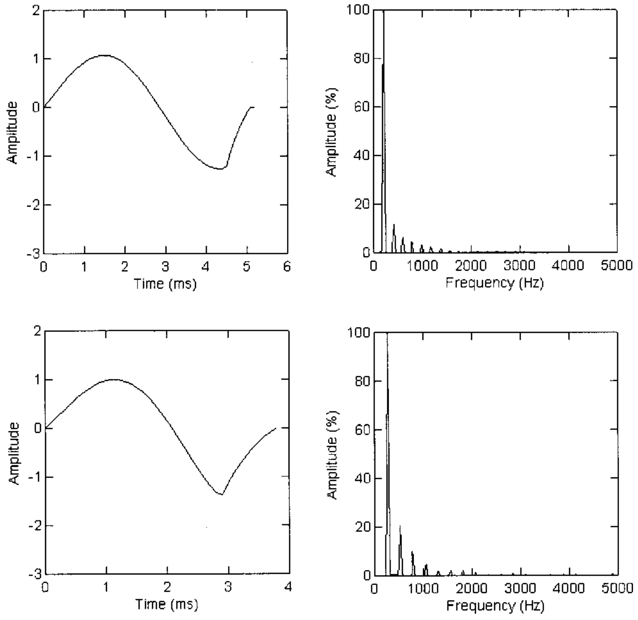

FIG. 2. Glottal volume velocity derivatives and associated linear source spectra for two voices for which listeners agreed well in their choice of signal-to-noise ratios.

fact that most of the responses were within a difference limen of each other.

Examination of estimated source spectra provides a preliminary explanation for the differences in listeners' sensitivity to changes in noise levels across stimuli. Figure 2 shows representative glottal volume velocity derivatives and the associated linear fast Fourier transform spectra for the synthetic stimuli about which listeners agreed precisely in their selection of noise levels. Figure 3 shows the volume velocity derivatives and source spectra for voices 5, 9, and 11, for which listeners varied in their chosen signal-to-noise ratio. The voices about which listeners agreed well in their noise settings all had sources with very little energy above the first or second harmonic. In contrast, voices 5, 9, and 11 all had sources with significant amounts of high-frequency harmonic energy, in addition to the excitation provided by the aperiodic component of the source [cf. Cranen and Schroeter (1995), who describe the different effects of posterior versus medial glottal gaps on source spectral slopes]. These two groups of stimuli differed significantly in values of the LF composite parameter RA (Fant and Lin, 1988), which measures the amount of harmonic source energy in the higher frequencies [$F(1,10)=75.35$, $p<0.05$]. Although any interpretation must be tentative due to the limited amount of data, small differences in noise levels are apparently more difficult to discriminate in the presence of harmonic energy in the higher frequencies, presumably due to masking effects. Unlike rating scale tasks, the analysis-synthesis protocol could be used in a straightforward manner to test this hypothesis (and others like it), and to derive the precise relationship between listener sensitivity and the characteristics of the harmonic and inharmonic components of the source.

IV. GENERAL DISCUSSION

These results indicate that listeners do in fact agree in their perceptual assessments of pathological voice quality, and that tools can be devised to measure perception reliably. Many previous studies have demonstrated disagreement among listeners in their use of traditional unidimensional rating scales for voice quality. However, the use of an analysis-synthesis protocol greatly simplifies the task of assessing quality. It does not require listeners to refer to varying internal standards for individual qualities, and facilitates attending to single acoustic dimensions in complex, varying contexts. Thus, this task appears to control several major sources of variability in listeners' responses, providing perceptual measures that are relatively free of the confounding influences of cognitive and task-related variables.

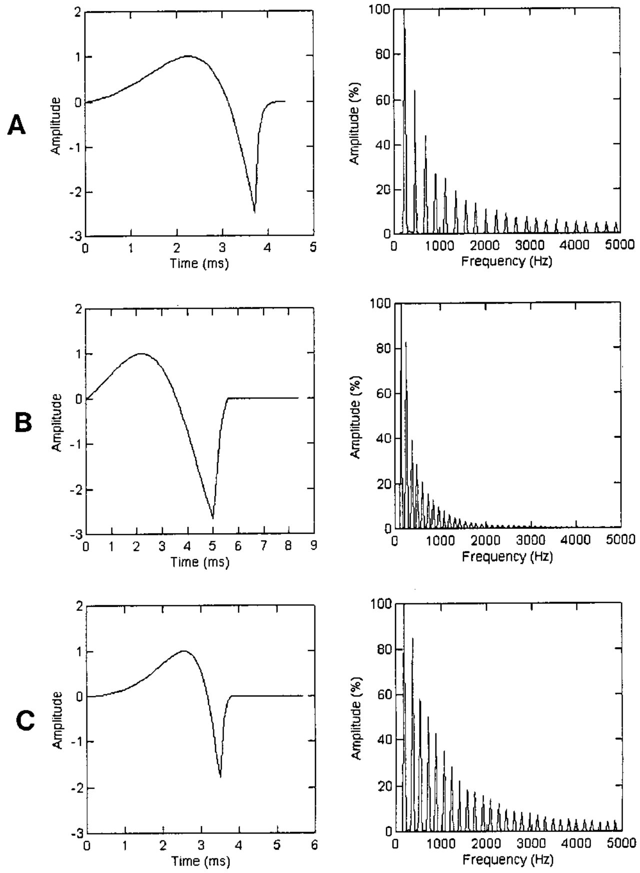

FIG. 3. Glottal volume velocity derivatives and associated linear source spectra for the three voices for which listeners gave varying responses when asked to match the signal-to-noise ratio. Voice numbers correspond to Table I: (a) voice 5, (b) voice 9, and (c) voice 11.

A reliable, valid protocol for quantifying voice quality perception would have many applications. Researchers and clinicians alike have long sought objective measures of voice quality, for example to examine prosodic or linguistic changes in voice quality, to model differences between speakers in individual voice quality, or to track improvement within a single speaker with treatment. One traditional research approach has been to gather scalar ratings for individual voice qualities and calculate correlations between average ratings and acoustic or other instrumental measures of the rated voices. Obviously, this method cannot establish a direct link between the perception of a specific voice and the acoustic element that gave rise to the perception, due to the limitations of correlational approaches and to the fact that listener disagreements undermine the interpretation of average ratings. The analysis-synthesis method allows listeners to record their perception of voice quality parametrically and objectively by selecting the level of a set of acoustic attributes. Because this approach uses speech synthesis, signals can be manipulated systematically to demonstrate direct causation between acoustic parameters and perceived quality, overcoming the first of these limitations; and the results presented here suggest that the listening task produces reliable listener responses, removing the second impediment to understanding the complex association between a signal and its perception. Much more research is certainly needed to determine a meaningful, parsimonious set of acoustic parameters that successfully characterizes all possible normal and pathological voice qualities. However, such a set could obviate the need for voice quality labels, allowing researchers and clinicians to replace quality labels with acoustic parameters whose levels objectively, completely, and validly specify the voice quality of interest.

ACKNOWLEDGMENTS

The research was supported by NIDCD Grant No. DC01797. Thanks are due to Melissa Epstein for creating stimuli and testing listeners, and to Norma Antonanzas and Brian Gabelman for extensive and creative programming support. We are also grateful to Keith Kluender, James Hillenbrand, and two anonymous reviewers for their thoughtful comments on a previous version of this article.

[1] This software is available at http://www.surgery.medsch.ucla.edu/glottalaffairs/software.

[2] A more recent version of this software allows $f0$ tracks to be generated and used instead of computed tremors, if desired.

[3] Exact values of the signal-to-noise ratio could also be typed into a small window next to the sliding cursor, but listeners in general did not use this option.

Cranen, B., and Schroeter, J. (**1995**). ''Modeling a leaky glottis,'' J. Phonetics **23**, 165–177.

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. (**1994**). ''Consistency and reliability of voice quality ratings for different types of speech fragments,'' J. Speech Hear. Res. **37**, 985–1000.

Egan, J. P. (**1975**). *Signal Detection Theory and ROC Analysis* (Academic, New York).

Epstein, M., Gabelman, B., Antoñanzas-Barroso, N., Gerratt, B., and Kreiman, J. (**1999**). ''Source model adequacy for pathological voice synthesis,'' in *Proc. ICPhS99*, pp. 2049–2052.

Fant, G., Liljencrants, J., and Lin, Q. (**1985**). ''A four-parameter model of glottal flow,'' Speech Transmission Lab. Quart. Prog. Status Rep. **4**, 1–13.

Fant, G., and Lin, Q. (**1988**). ''Frequency domain interpretation and derivation of glottal flow parameters,'' STL-QPSR **2-3**, 1–21.

Gelfer, M. P. (**1988**). ''Perceptual attributes of voice: Development and use of rating scales,'' J. Voice **2**, 320–326.

Gerratt, B. R., Kreiman, J., Antoñanzas-Barroso, N., and Berke, G. S. (**1993**). ''Comparing internal and external standards in voice quality judgments,'' J. Speech Hear. Res. **36**, 14–20.

Gescheider, G. A. (**1997**). *Psychophysics: The Fundamentals*, 3rd ed. (Erlbaum, Mahwah, NJ).

Gescheider, G. A., and Hughson, B. A. (**1991**). ''Stimulus context and absolute magnitude estimation: A study of individual differences,'' Percept. Psychophys. **50**, 45–57.

Hillenbrand, J., Cleveland, R., and Erickson, R. (**1994**). ''Acoustic correlates of breathy vocal quality,'' J. Speech Hear. Res. **37**, 769–778.

Hillenbrand, J., and Houde, R. A. (**1996**). ''Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech,'' J. Speech Hear. Res. **39**, 311–321.

Hirano, M., Hibi, S., Yoshida, T., Hirade, Y., Kasuya, H., and Kikuchi, Y. (**1988**). ''Acoustic analysis of pathological voice,'' Acta Oto-Laryngol. **105**, 432–438.

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.

Kreiman, J., Gabelman, B., and Gerratt, B. R. (**2001**). ''Perceptual and acoustic modeling of vocal tremor,'' submitted for publication.

Kreiman, J., and Gerratt, B. R. (**1998**). ''Validity of rating scale measures of voice quality,'' J. Acoust. Soc. Am. **104**, 1598–1608.

Kreiman, J., and Gerratt, B. R. (**2000**). ''Sources of listener disagreement in voice quality assessment,'' J. Acoust. Soc. Am. **108**, 1867–1879.

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., Kempster, G., Erman, A., and Berke, G. S. (**1993**). ''Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research,'' J. Speech Hear. Res. **36**, 21–40.

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

Michaelis, D., Frohlich, M., and Strube, H. (**1998**). ''Selection and combination of acoustic features for the description of pathologic voices,'' J. Acoust. Soc. Am. **103**, 1628–1639.

Orlikoff, R. O. (**1999**). ''The perceived role of voice perception in clinical practice,'' Phonoscope **2**, 87–106.

Poulton, E. C. (**1979**). ''Models for biases in judging sensory magnitude,'' Psychol. Bull. **86**, 777–803.

Sundberg, J. (**1987**). *The Science of the Singing Voice* (Northern Illinois U.P., De Kalb, IL).

Swets, J. A., and Pickett, R. M. (**1982**). *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory* (Academic, New York).

Verdonck-de Leeuw, I. M. (**1998**). ''Perceptual analysis of voice quality: Trained and naive raters, and self-ratings,'' in *Proceedings of Voicedata98 Symposium on Databases in Voice Quality Research and Education,* edited by G. de Krom (Utrecht Institute of Linguistics OTS, Utrecht), pp. 12–15.

Wedell, D. H., Parducci, A., and Lane, M. (**1990**). ''Reducing the dependence of clinical judgment on the immediate context: Effects of number of categories and type of anchor,'' J. Pers. Soc. Psychol. **58**, 319–329.

Weismer, G., and Liss, J. (**1991**). ''Reductionism is a dead-end in speech research: Perspectives on a new direction,'' in *Dysarthria and Apraxia of Speech: Perspectives on Management,* edited by K. Yorkston, C. Moore, and D. Beukelman (Brookes, Baltimore), pp. 15–27.

Wuyts, F. L., DeBodt, M. S., and Van de Heyning, P. H. (**1999**). ''Is the reliability of a visual analog scale higher than an ordinal scale? An experiment with the GRBAS scale for the perceptual evaluation of dysphonia,'' J. Voice **13**, 508–517.

Yumoto, E., Gould, W. J., and Baer, T. (**1982**). ''Harmonics-to-noise ratio as an index of the degree of hoarseness,'' J. Acoust. Soc. Am. **71**, 1544–1550.