When and why listeners disagree in voice quality assessment tasks[a)]

Jody Kreiman,[b)] Bruce R. Gerratt, and Mika Ito
Division of Head and Neck Surgery, UCLA School of Medicine, 31–24 Rehab Center, Los Angeles, California 90095

(Received 13 December 2006; revised 18 July 2007; accepted 19 July 2007)

Modeling sources of listener variability in voice quality assessment is the first step in developing reliable, valid protocols for measuring quality, and provides insight into the reasons that listeners disagree in their quality assessments. This study examined the adequacy of one such model by quantifying the contributions of four factors to interrater variability: instability of listeners' internal standards for different qualities, difficulties isolating individual attributes in voice patterns, scale resolution, and the magnitude of the attribute being measured. One hundred twenty listeners in six experiments assessed vocal quality in tasks that differed in scale resolution, in the presence/absence of comparison stimuli, and in the extent to which the comparison stimuli (if present) matched the target voices. These factors accounted for 84.2% of the variance in the likelihood that listeners would agree exactly in their assessments. Providing listeners with comparison stimuli that matched the target voices doubled the likelihood that they would agree exactly. Listeners also agreed significantly better when assessing quality on continuous versus six-point scales. These results indicate that interrater variability is an issue of task design, not of listener unreliability.
 [DOI: 10.1121/1.2770547]

PACS number(s): 43.71.Bp [ARB] Pages: 2354–2364

I. INTRODUCTION

Previous research has shown that listeners often disagree with one another in their ratings of voice quality (e.g., Cullinan *et al.*, 1963; Ludlow, 1981; Webb *et al.*, 2003; see Kreiman *et al.*, 1993, for review), and suggests that these disagreements are the result of the methods used to gather ratings and not of flawed or inconsistent perceptual abilities (e.g., Gerratt and Kreiman, 2001; Kreiman and Gerratt, 2005). Ideally, measurement protocols for assessing voice quality should control all the known sources of irrelevant or unwanted variability in listeners' responses. For this reason, modeling the different sources of listener variability in voice quality assessment is the first step in developing accurate, reliable, and valid protocols for measuring quality. Such a model could also provide insight into the perceptual processes listeners use when assessing quality. This study examines the adequacy of one model of the sources of variability in listeners' judgments of voice quality, by quantifying the contributions of four factors to the total variance in a set of quality judgments: the instability of listeners' internal standards for different qualities, difficulties isolating individual attributes in complex acoustic voice patterns, measurement scale resolution, and the magnitude of the attribute being measured.

Previous research has documented the importance of these four factors as sources of variability in quality assessments. We have claimed that traditional rating scale protocols require listeners to judge voices by comparing the vocal stimulus to a standard held in memory for the attribute in question. Listeners' internal standards are idiosyncratic, and vary both within and across listeners with a number of factors, including listeners' previous experiences with voices and the context in which judgments are made (Kreiman *et al.*, 1990; Verdonck-de Leeuw, 1998). When listeners are provided with external comparison stimuli as a referent for their judgments, dependence on varying internal standards is eliminated and rater agreement increases (Gerratt *et al.*, 1993; Gerratt and Kreiman, 2001). This comparison stimulus effect forms the first factor in this experiment. The second factor represents difficulty isolating the dimension to be judged in a complex, time-varying vocal pattern. This factor contributes variability in traditional rating scale tasks, and also in tasks where listeners judge quality with respect to comparison stimuli or anchors. In tasks with comparison stimuli, listeners try to compare magnitudes of the quality being studied in the differing acoustic contexts provided by the test and comparison stimuli. When the comparison stimulus is matched to the test stimulus, so that they differ only in the aspect being judged, listener agreement increases, because the dimension of interest moves to the perceptual foreground as listeners compare magnitudes in otherwise matching acoustic patterns. This helps listeners isolate and focus attention on that dimension (Kreiman and Gerratt, 2005). The third factor contributing to variability—scale resolution—measures the error that occurs when listeners have difficulty selecting a response because the quality of a test stimulus falls between two scale values or two comparison stimuli, or when scale steps or comparison stimuli are so close together that listeners cannot discriminate between ad-

[a)] Portions of this work were presented at the 150th Meeting of the Acoustical Society of America and at the 4th Joint Meeting of the Acoustical Society of America and the Acoustical Society of Japan.
[b)] Electronic mail: jkreiman@ucla.edu

TABLE I. Experimental Conditions

	Continuous scale	Six-point scale
Comparison stimulus condition (acoustic context)		
Matched exactly to each target voice (*custom* condition)	Experiment 1	Experiment 2
Matched to the target voices with respect to sex only (*generic* condition)	Experiment 3	Experiment 4
No comparison stimulus	Experiment 5	Experiment 6

jacent levels, and consequently must choose between perceptually equivalent response alternatives (Gerratt *et al.*, 1993). The final factor—the magnitude of the attribute being rated—quantifies the systematic variability in listener agreement levels as a function of the mean rating for a voice. This factor incorporates several previous findings relating agreement levels to the characteristics of the specific voices under study. For example, agreement is best near the endpoints of the scale and worse near the midpoint when listeners rate voices on traditional scales for individual qualities (Kreiman and Gerratt, 1998), in part because the acoustic characteristics of some voices correspond to the limiting case of the quality in question (e.g., aphonia is the greatest possible extent of breathiness; Kreiman and Gerratt, 2005) and in part because fewer response alternatives are available at the end of a scale. In contrast, in method of adjustment tasks (in which listeners adjust the parameters of a synthetic stimulus until it matches the target), agreement increases with the mean rating for a voice, presumably reflecting the difficulty of assessing an acoustic characteristic (and the relatively large difference limens) near the threshold of detection (Kreiman and Gerratt, 2005; Shrivastav and Sapienza, 2003; 2006). By modeling these effects, which depend on the specific set of voices under study, this factor serves to tune the model to the individual data set.

Although these previous studies have documented the individual contributions of these factors to rating variability, no model has examined their interactions, quantified the extent to which as a group they adequately account for interrater variability, or determined the size of their relative contributions to overall variability. To this end, six experiments were devised to examine the effects of these factors on listener agreement in judgments of breathiness and noise-to-signal ratios (NSRs) for pathological voices (Table I). We selected these attributes for study because breathiness is often defined in terms of spectral noise levels, and perceived breathiness has been repeatedly associated with the presence of turbulent noise (e.g., Hillenbrand *et al.*, 1994; Klatt and Klatt, 1990; Shrivastav and Sapienza, 2003; Yiu and Ng, 2004). This definitional equivalence between noise levels and breathiness made it reasonable to compare data from experiments in which listeners manipulated an acoustic attribute (the NSR) and experiments in which they rated a perceptual characteristic (breathiness). Listeners are also perceptually sensitive to spectral noise levels (Gerratt and Kreiman, 2001; Kreiman and Gerratt, 2005), so any differences observed between experimental conditions could unambiguously be assigned to the task, and not to difficulties hearing the quality being measured.

The first experimental factor—the presence or absence of an external comparison stimulus during the voice assessment task—examined the importance of unstable internal standards for different levels of a vocal attribute. In Experiments 1, 2, 3, and 4, listeners were provided with comparison stimuli; in Experiments 5 and 6 they made judgments solely on the basis of their internal standards. The second factor assessed the effect of listeners' (in)abilities to separate the relevant dimension perceptually from a complex pattern, by providing synthesized comparison stimuli that matched the test voices exactly in quality (hereafter referred to as "custom comparison stimuli;" Experiments 1 and 2) or matched them only with respect to the speaker's sex ("generic comparison stimuli;" Experiments 3 and 4). The third factor—scale resolution—was assessed by comparing ratings on responses from a continuous, visual-analog-type scale (Experiments 1, 3, and 5) with a six-point equal-appearing interval-type rating scale (Experiments 2, 4, and 6). The final experimental factor—the distance of a voice from the extreme values of an attribute—was quantified as the overall mean rating for each voice on each scale. This value was included as a covariate in each of these experiments, to assess the contribution of this fourth factor to rating variability. We hypothesized that the highest levels of interrater agreement would occur in Experiment 1 (in which all hypothetical sources of interrater variability were controlled) and the worst agreement in Experiment 6 (nothing controlled).

II. METHOD

A. Voice samples

Forty voices (20 male speakers, 20 female speakers) were selected at random from a library of pathological voices recorded under identical conditions. No attempt was made to select stimuli that possessed any particular quality, nor did we attempt to create a continuum from mild to severe breathiness, although voices did span the range from near-normal to severely deviant in terms of overall level of pathology. Pathological voices are appealing stimuli not only because of their clinical interest, but also because they encompass the full range of quality that can be produced by a human voice. A 1 s sample was excerpted from the middle of a sustained /a/. Vowels were studied (rather than continuous speech) because steady-state vowels are routinely used for evaluating pathological voice quality and carry much information about the voice source; because analysis and synthesis are far more straightforward than for continuous speech; and because the simpler acoustic structure of steady-state vowels typically yields responses from listeners reflecting simpler perceptual strategies that are more easily interpreted.

Each sample was copied using a custom formant synthesizer optimized for precisely modeling pathologic voice quality. Analysis and synthesis procedures are described in detail elsewhere (Kreiman *et al.*, 2006). Briefly, the synthesizer sampling rate was fixed at 10 kHz. Parameters describing the harmonic part of the voice source were estimated by

inverse filtering a representative cycle of phonation for each voice using the method described by Javkin *et al.* (1987). The extracted pulses were least-squares fit with a modified Liljencrants–Fant (LF) source model (Fant *et al.*, 1985; Kreiman *et al.*, 2007), and the LF model parameters were used to specify the harmonic voice source in the synthesizer. The spectral characteristics of the inharmonic part of the source (the noise excitation) were estimated using a cepstral-domain analysis similar to that described by de Krom (1993). Spectrally shaped noise was synthesized by passing white noise through a 100 tap finite impulse response filter fitted to that noise spectrum. To model the F0 contour, F0 was tracked pulse by pulse on the time domain waveform by an automatic algorithm, and a train of LF pulses with the appropriate periods was added to the noise time series to create a complete glottal source waveform. Formant frequencies and bandwidths were estimated using autocorrelation linear predictive coding analysis with a window of 25.6 ms (increased to 51.2 ms when F0 was near or below 100 Hz). The complete synthesized source was filtered through the vocal tract model to generate a preliminary version of the synthetic voice.

B. Listening pretest

Measurement of many acoustic parameters is difficult and often inaccurate when phonation departs from periodicity (e.g., Titze, 1994; Bielamowicz *et al.*, 1996). For this reason, all synthesizer parameters were perceptually adjusted from their measured values by the first author until the synthetic copies perfectly matched the natural target voices. A pretest was used to verify the accuracy of the adjustments and the synthesis. Twelve listeners [UCLA students and staff; 21–55 years of age; mean age=37.9 years; standard deviation (sd)=12.6 years] heard pairs of voices. On half the trials, a synthetic voice sample was paired with its natural counterpart, and on the other half, stimuli were identical. Each pair was repeated twice, for a total of 160 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 five-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. Stimuli were presented in free field at a comfortable constant listening level. Testing lasted approximately 30 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 79.2% to 100%, with an average of 95.3% (sd=5.1%). False alarm rates ranged from 41.7% to 100%, with an average of 73.1% (sd =14.5%). Same/different responses for each voice were then combined with confidence ratings to create a ten-point scale ranging from "positive voices are the same" (1) to "positive voices are different" (10). Values of d' were calculated from these data for each voice; values ranged from 0.002 to 0.97 (mean $d'=0.43$; sd=0.274). Receiver operating characteristics (ROCs) consisting of nine points each were also constructed from these recoded data following the procedure described by Green and Swets (1966; see also Macmillan and Creelman, 2005). 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 indicate that listeners were unable to consistently distinguish the synthetic copies from the natural samples. We conclude that the synthetic tokens provide good models of the quality of the natural target voice samples.

C. Comparison stimuli

Four sets of synthetic comparison stimuli were constructed. The first set (*custom comparison stimuli*) comprised 40 synthetic voice tokens modeled on the 40 natural voice samples described earlier. Each of these 40 stimuli was synthesized using the synthesis parameters that produced perceptually exact copies of the voices, as described previously. Listeners manipulated the NSR of these synthetic voices in Experiment 1, as described in the following.

The second set of custom comparison stimuli comprised six additional versions of each synthetic voice, created for use in Experiment 2 (see Table I). The six versions differed only in NSR levels, in steps. All custom comparison stimuli were 1 s in duration. The first stimulus in each of the 40 series was created with NSR equal to −50 dB (noise-free), and the second used the lowest NSR value at which noise was consistently detectable for that voice, as determined by pilot study. This value ranged from −36 to −17 dB, with a mean of −23.8 dB (sd=4.08 dB). The sixth stimulus in each series was created with NSR equal to 0 dB, and the remaining three stimuli evenly spanned the acoustic range between the second and sixth stimuli. The change in NSR between these last four stimuli ranged from 4.25 to 9 dB, with a mean interstimulus step size of 5.94 dB (sd=1.02 dB). Pilot experiments confirmed that all comparison stimuli in each series were easily discriminable from their immediate neighbors in the series.

The third set of comparison stimuli (*generic comparison stimuli*) comprised two synthetic voices, one male and one female. Synthesis parameters for these voices are listed in Table II. Both voices were created using the same LF source pulse shape and F0 contour shape, although mean F0 differed for the male and female tokens (female F0=228 Hz, male F0=135 Hz). Because listeners' sensitivity to changes in the NSR depends, in part, on the shape of the harmonic spectrum (Gerratt and Kreiman, 2001; Kreiman and Gerratt, 2005), these voices were synthesized using a source with a moderate amount of high frequency harmonic energy (neither sinusoidal nor impulse-like; Fig. 1). This resulted in a scale for which small changes in NSR were perceptible, and for which the resulting stimuli sounded natural across NSR values. Vocal tract parameters were modeled on natural male and female voices that were not used elsewhere in these experiments. Listeners manipulated the NSR of these synthetic

TABLE II. Synthesis parameters for generic comparison stimuli (Experiments 3 and 4) [a]

	Female voice	Male voice
Source parameters		
F0 (Hz)	228	135
Tremor rate (Hz)	4	4
Tremor amplitude (Hz)	0.8	0.8
Parameters of the modified LF model:		
t_p	0.002917	0.004063
t_e	0.003866	0.005385
Ee	1.651	1.651
t_2	0.004247	0.005916
t_c	0.004819	0.006712
Vocal tract parameters		
F1/B1 (Hz/Hz)	670/151	792/96
F2/B2 (Hz/Hz)	1128/134	1224/399
F3/B3 (Hz/Hz)	2337/143	2011/557
F4/B4 (Hz/Hz)	3101/263	3027/119
F5/B5 (Hz/Hz)	4200/500	3942/402
NSR values (dB)	−50	−50
	−25	−28
	−18.75	−21
	−12.5	−14
	−6.25	−7
	0	0

[a]These synthetic voices were modeled after natural vowels produced by two individuals. The first two formants of the man's vowel were higher than those for the woman's, but the voices are unambiguously male and female despite this coincidental occurrence.

voices in Experiment 3, as described in the following.

The fourth set of comparison stimuli included six additional versions of these two voices, which were created for use in Experiment 4. As for the custom comparison stimuli, the six versions differed only in NSR levels, in steps. The first stimulus in each six-step series was synthesized with NSR = −50 dB (noise-free). The second stimulus was created with the minimum NSR value at which noise was clearly perceptible in the voice (determined in pilot studies; −25 dB for the female voice, and −28 dB for the male voice). The sixth stimulus had an NSR of 0 dB (extremely noisy). The remaining three stimuli evenly spanned the acoustic range between the second and sixth stimuli. All stimuli were 1 s in duration, and pilot experiments again confirmed that all stimuli in the series were easily discriminable from their neighbors, for both voices.

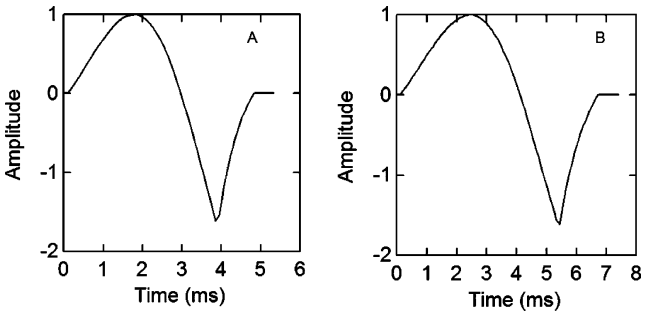

FIG. 1. Source pulses used in synthesis of the generic comparison stimuli in Experiments 3 and 4. (A) Female and (B) male voice sources

D. Listeners and testing environment

A total of 120 listeners (UCLA students and staff members) participated in these experiments, 20 in each task. No listener participated in more than one task, and none had participated in any previous experiments involving these voice stimuli. All listeners reported normal hearing. They ranged in age from 18 to 60 (mean=28.7, sd=10.23).

All six experiments took place in a double-walled sound suite. Stimuli were presented in free field at a comfortable listening level. Listeners practiced with the software using additional voices until they were comfortable and understood the task, but they did not hear the entire stimulus set prior to beginning the experiments. However, recent data (Gerratt and Kreiman, 2007) indicate that listeners who hear voices repeatedly do not adjust the range of their ratings as they learn the stimulus set, but instead remember the score given to each stimulus. This suggests that (lack of) practice with the stimuli did not contribute significantly to the present results. Listeners controlled the pace of the experiments, and were able to play the test and comparison voices as often as they wished, in any order, prior to responding.

E. Experiments 1 and 3: Voice quality assessments using continuous scales and comparison stimuli

In Experiments 1 and 3, listeners heard the 40 test stimuli one at a time in a unique random order. Each natural test voice was paired with a synthetic comparison sample (custom matched to the test voice in Experiment 1, and generic in Experiment 3). For each target voice in each experiment, listeners manipulated the NSR of the comparison stimulus in a method of adjustment task by adjusting a sliding cursor with a mouse until they determined that the synthetic noise level matched that of the natural test voice as closely as possible. The noise scale ranged from −50 dB (noise-free) to 0 dB (extremely noisy), and was set to −50 dB at the beginning of each trial. Listeners were able to play both the test and comparison stimuli in any order as often as they desired as they made their adjustments. In Experiment 1, listeners always judged the test and comparison noise levels in the same acoustic context, because each comparison stimulus matched the corresponding test voice (except for the initial NSR level). Because Experiment 3 used generic comparison stimuli, listeners were required to compare noise levels in unmatched acoustic contexts. Because listeners worked interactively with these voices, no trials were repeated due to the likelihood of remembering individual tokens. Testing lasted about 60 min, on average.

F. Experiments 2 and 4: Voice quality assessments using six point scales and comparison stimuli

Listeners in Experiments 2 and 4 heard the 40 natural test voices one at a time in a unique random order. In each trial, they compared the noise level of the natural test stimulus to those of the tokens in the series of six synthetic comparison stimuli, and decided which of the comparison stimuli best matched the test voice with respect to the amount of noise present. In Experiment 2 listeners compared the test

voices to custom comparison stimuli; Experiment 4 used generic comparison stimuli. In both experiments, listeners were able to replay the test and comparison stimuli as often as they wished, in any order, before responding. Eight randomly selected trials (20%) were repeated to assess test-retest reliability. Testing lasted approximately 25 min on average.

G. Experiments 5 and 6: Voice quality assessments without comparison stimuli

In Experiments 5 and 6, listeners rated the breathiness of each of the 40 test stimuli on traditional rating scales without comparison stimuli. In Experiment 5, they rated the voices by moving a sliding cursor along a line with a mouse (a visual analog-type scale, as in Experiments 1 and 3). Responses were stored as percentages of the scale length from 0 (noise-free) to 100 (extremely noisy). In Experiment 6, listeners rated the breathiness of each stimulus by clicking a number between 1and 6 on a computer display (an equal-appearing interval-type scale, as in Experiments 2 and 4). In both experiments they heard the test stimuli one at a time, in a unique random order. Eight randomly selected stimuli (20%) were repeated to assess test–retest reliability, for a total of 48 trials/listener. Testing took an average of about 15 min to complete.

H. Calculating agreement levels for the different tasks

To assess interlistener agreement, we calculated the likelihood that two randomly selected listeners would agree exactly in their response for a given voice, for each of the 40 voices in each experiment. This approach provides information about systematic patterns of agreement among listeners for specific ranges of rating scales, and contrasts with more commonly used measures like Cronbach's alpha and intraclass correlations that sum across voices to provide a single measure of rater concordance (Kreiman and Gerratt, 1998).

Although no method of calculating agreement can guarantee that listeners mean the same thing when they produce the same response, even within a single protocol, the following procedures were used to equilibrate the response scales in the six experiments as nearly as possible so that agreement levels could be fairly compared across tasks. For tasks using the six-point rating scale, the probability that two listeners would agree exactly was calculated for each voice by examining the difference in ratings between all possible pairs of raters (190 comparisons/voice) and calculating the percentage of exact agreements out of a possible total of 190. For tasks using continuous scales with comparison stimuli, two NSR responses were considered to agree exactly when the difference between them was less than the NSR difference between adjacent comparison stimuli on the six-point scale for that voice. The probability of exact agreement between two raters was calculated based on these quantized ratings in the same manner as it was for the six-point scale protocols. This quantization procedure served to equilibrate the continuous and six-point scales and ensured that responses treated as different were in fact perceptually discriminable. At the same time, this procedure re-centers the window of agreement around each individual response, rather than locking it to the fixed intervals of the six-point scale. Because agreement was not referenced to arbitrary scale divisions, responses that were close together were always considered to agree, regardless of where they fell on the scale, consistent with the continuous nature of these scales.

For ratings on a continuous scale without comparison stimuli (Experiment 5), we defined "exact agreement" as agreement within 20 units on a 100 unit linear scale (five intervals, equivalent to six points). This second quantization procedure had the effect of creating a scale whose intervals are equal in width to those of the six-point scales, but that are again centered around the individual ratings on the continuous scales. As mentioned earlier, responses that were close together were always considered to agree, regardless of where they fell on the scale. The probability of exact agreement was calculated based on these quantized ratings, as described previously.

TABLE III. Test-retest agreement

	Exact agreement (%)	Mean difference between first and second ratings (sd)
Protocol		
Six-point scale Custom comparison stimuli (Experiment 2)	70.6	0.31 (0.50)
Six-point scale Generic comparison stimuli (Experiment 4)	49.4	0.77 (0.91)
Continuous scale No comparison stimuli (Experiment 5)	66.3	0.52 (0.88)
Six point scale No comparison stimuli (Experiment 6)	50.6	0.68 (0.83)

III. RESULTS

A. Test–retest agreement

Test–retest agreement levels for rating protocols without comparison stimuli (Experiments 5 and 6) and for experiments using six-point rating scales with comparison stimuli (Experiments 2 and 4) are given in Table III. As described previously, test–retest agreement was not assessed for the method-of-adjustment tasks using continuous scales and comparison stimuli (Experiments 1 and 3), due to the probability of learning effects. Listeners were reasonably self-consistent in all these experiments, but for custom comparison stimuli (Experiment 2) the mean difference between the first and second rating was significantly lower than in any other condition [so that ratings agreed more closely; $F(1,636)=9.98$, $p<0.05$; Scheffé post-hoc comparisons, $p<0.05$]. Mean differences between repeated ratings were statistically indistinguishable when listeners made ratings on the six-point scale without comparison stimuli (Experiment 6) and when they made similar ratings with generic comparison stimuli (Experiment 4; $p>0.05$). Test–retest agreement was slightly but significantly higher for ratings on the con-

TABLE IV. Mean probability of exact agreement between two listeners. Standard deviations are given parenthetically

	Scale resolution	
	Continuous scale	Six-point scale
Comparison stimulus condition (acoustic context)		
Custom comparison stimuli	0.96 (0.12)	0.63 (0.13)
Generic comparison stimuli	0.42 (0.11)	0.28 (0.08)
No comparison stimuli	0.53 (0.14)	0.30 (0.14)

tinuous scale without comparison stimuli than for the six-point scale ratings with generic comparison stimuli ($p < 0.05$).

B. Between-task differences in response variability

Analysis of covariance (ANCOVA) indicated that comparison stimulus condition (custom, generic, and none), scale resolution (continuous versus six point), and the mean rating for each voice all had significant effects on the probability that two listeners would agree exactly. The independent variables and covariate (mean rating) in this experiment together accounted for 84.2% of the variance in agreement levels. The mean probabilities of exact agreement between two raters are given for each task in Table IV.

Across tasks, the probability of agreement covaried significantly with the mean rating [$F(1,233)=55.50$, $p<0.05$, 4.3% variance accounted for]. Figure 2 shows the probability of exact listener agreement for the six experiments, plotted as a function of the mean rating for each voice in each task. The

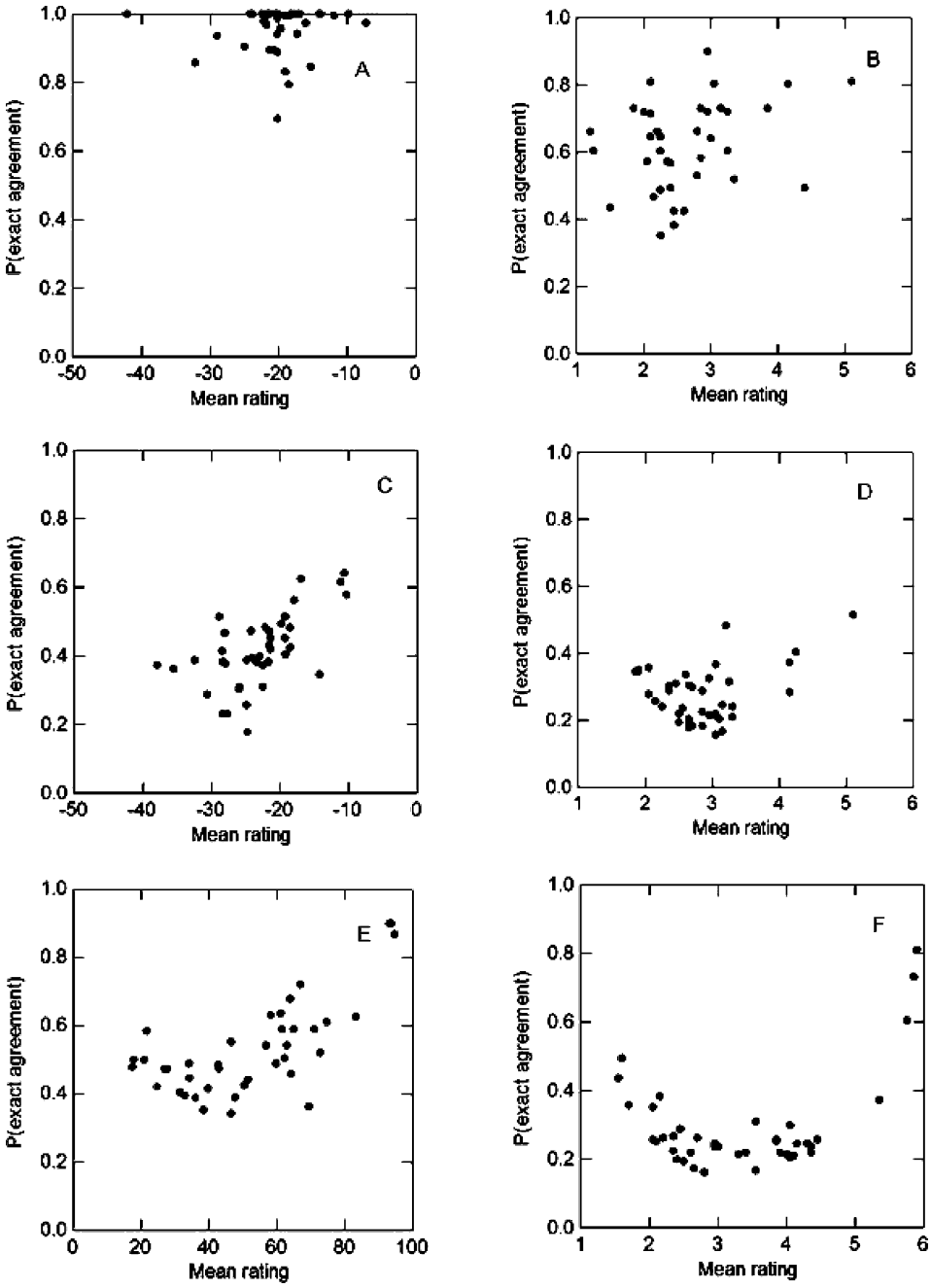

FIG. 2. The probability that two listeners would agree exactly in their rating, versus the mean rating for each voice sample for the six experimental tasks. (A) Continuous scale, custom comparison stimulus; (B) six-point scale, custom comparison stimulus; (C) continuous scale, generic comparison stimulus; (D) six-point scale, generic comparison stimulus; (E) continuous scale, no comparison stimulus; and (F) six-point scale, no comparison stimulus.

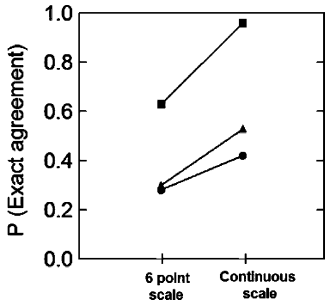

FIG. 3. Interaction between comparison stimulus condition and scale resolution in the ANCOVA analysis. Squares represent custom comparison stimuli; circles represent generic comparison stimuli; and triangles represent the conditions without comparison stimuli.

precise patterns of agreement differed somewhat from task to task. The pattern seen in panels E and F for the traditional rating tasks without comparison stimuli agrees with that usually seen in such rating scale experiments, in which listeners agree best on stimuli at the extremes of the scale and worst in the middle range (Kreiman and Gerratt, 1998). Quadratic equations fit the data in these panels substantially better than did linear models, consistent with this interpretation (panel E: r^2 for the nonlinear model=0.69, versus 0.46 for the linear model; panel F: r^2 for the nonlinear model=0.86, versus 0.14 for the linear model). A similar pattern is apparent in panel D (six-point scale, generic comparison stimuli; r^2 for the nonlinear model=0.41, versus 0.06 for the linear model). In contrast, agreement increased as a linear function of the mean rating for the task using generic comparison stimuli and a continuous scale, although the difference in fit was less pronounced than for panels E and F (panel C; r^2 for the linear model=0.59, versus 0.43 for the nonlinear model). No significant linear or nonlinear relationship was found between the mean rating and agreement levels for the tasks using custom comparison stimuli (panels A and B). (A significance level of $p<0.05$ was applied for all analyses).

The main effect of comparison stimulus condition accounted for 64.3% of the variance in agreement levels [$F(2,233)=441.58$, $p<0.05$]. The probability of exact agreement was substantially greater when listeners heard custom comparison stimuli than for the other two conditions [mean =0.79, versus mean=0.35 for generic comparison stimuli, and mean=0.41 when no comparison stimulus was present; planned comparisons, $F(1,233)=651.61$, $p<0.05$]. The smaller difference between the other two comparison stimulus conditions was also statistically significant: Listeners agreed significantly better when no comparison stimulus was present than they did for generic comparison stimuli [planned comparison, $F(1,233)=4.57$, $p<0.05$].

A significant main effect of scale resolution was also observed [$F(1,233)=176.73$, $p<0.05$; 13.0% variance accounted for], with better agreement occurring overall for the continuous scale (mean=0.63) than for the six-point scale (mean=0.40). Differences between scales were significant for all three comparison stimulus conditions, but the advantage provided by a continuous scale was relatively small when generic comparison stimuli were used, leading to a significant interaction effect [Fig. 3; $F(2,233)=115.90$, $p<0.05$; 2.6% variance accounted for]. This interaction effect appears to have its origins in response biases that inhere in tasks using six point scales with comparison stimuli (Experiments 2 and 4). Listeners in these tasks produced ratings that were significantly lower overall than those in the other four experiments, which did not differ [listener responses standardized to a fraction of the scale length prior to comparison; $F(5,234)=18.11$, $p<0.05$; Scheffé post-hoc comparisons, $p<0.05$]. (This effect can be seen in Fig. 2 as a shift to the left for the bulk of the data in panels B and D.) This suggests that when choosing between a comparison stimulus that contains too little noise and one that contains too much noise, listeners consistently selected the less-noisy alternative as the best match, so that mean ratings consistently underestimate the level of noise present in the stimulus relative to the other experimental conditions. This systematic response bias added error to the rating score itself, but at the same time it increased listener agreement by promoting a homogeneous strategy among listeners. In the case of the six-point scale/generic comparison stimuli task (Experiment 4), this boost in agreement levels has the net effect of minimizing the difference between that task and the condition using a 6 point scale without comparison stimuli (Experiment 6), as shown in Fig. 3.

The model coefficients generated by the ANCOVA allowed us to quantify the specific contribution each model component makes to listener agreement levels. Coefficients for all effects other than the covariate are shown in Table V, and the equation relating the dependent and independent variables is given in the Appendix . For ease of exposition, values in Table V have been converted to percentage change from mean levels in agreement. The constant in the model

TABLE V. The effects of the experimental variables on predicted levels of exact listener agreement. Each entry shows the values added to or subtracted from the model constant (36.44%) to calculate the likelihood of listener agreement for that experimental condition. The first value in each entry is the main effect of scale resolution; the second value is the main effect of comparison stimulus condition; the third value represents the interaction effect; and the fourth value is the sum of the first three. Adding the covariate correction of 0.3149 times the mean rating for that voice to these values produces the predicted agreement level for each voice in the experiment.

	Continuous scale	Six point scale
Comparison stimulus condition		
Custom comparison stimuli	+9.6+28.1+2.68=40.38	−9.6+28.1−2.68=15.82
Generic comparison stimuli	+9.6−16.1−5.35=−11.85	−9.6−16.1+5.35=−20.35
No comparison stimuli	+9.6−12.0+2.67=0.27	−9.6−12.0−2.67=−24.27

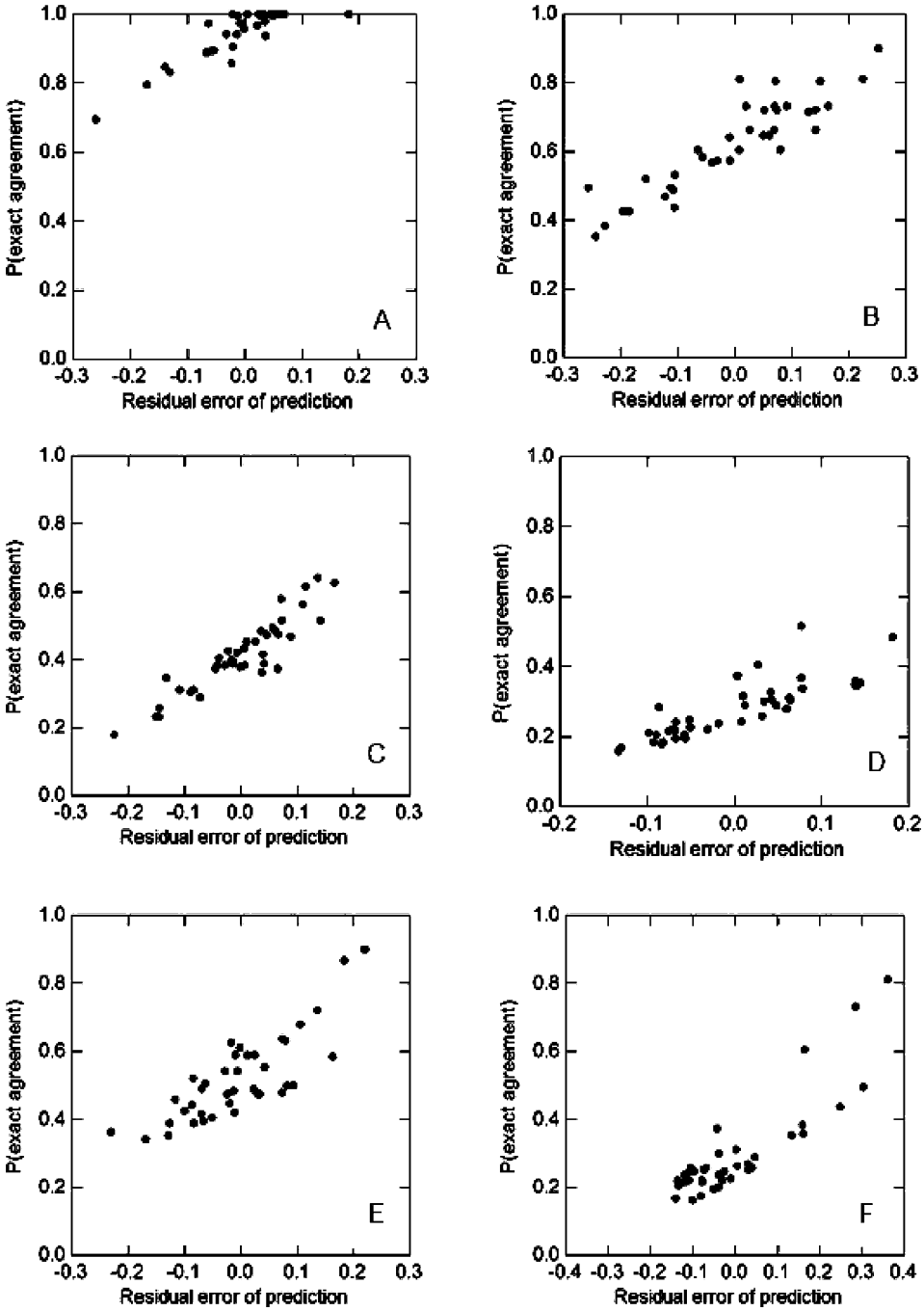

FIG. 4. The probability that two listeners would agree exactly in their ratings, versus residual errors of prediction for the six experimental tasks. (A) Continuous scale, custom comparison stimuli; (B) six-point scale, custom comparison stimuli; (C) continuous scale, generic comparison stimuli; (D) six-point scale, generic comparison stimuli; (E) continuous scale, no comparison stimuli; and (F) six-point scale, no comparison stimuli.

equals 36.44%, and represents the average likelihood across all voices and experimental conditions that two listeners will agree exactly. The main effect of comparison stimulus condition produced an increase from this average value of 28.1% in the probability of listener agreement when comparison stimuli matched the test stimuli, a decrease of 16.1% when comparison stimuli did not match test stimuli, and a decrease of 12.0% when no comparison stimuli were available. As a result of the main effect of scale resolution, the probability of exact listener agreement increased from average levels by 9.6% when the scale was continuous, and decreased by the same amount for a 6 point scale. The covariate effect applies equally to values in all cells in the design, and equals 0.3149 times the mean rating (standardized as a fraction of scale length) for that particular voice. The increase in predicted reliability due to this covariate thus ranges from a minimum of 0 at the left end of the scale, to a maximum of 31.49% at the right extreme.

Although overall this model fits the combined data well (R^2=0.842), fit did vary from task to task. Figure 4 shows for each task the difference between the observed and predicted agreement levels for each voice (residual error) plotted against the probability of exact agreement among raters. For Experiments 1–4 (panels A–D), these plots are unexceptional, with the data more or less symmetrically distributed about 0 error. Different patterns are apparent for the tasks without comparison stimuli (Experiments 5 and 6; panels E and F). For the six-point scale without comparison stimuli (panel F), the pattern of residuals indicates that the statistical model underestimates listener agreement near scale end-

points, yielding positive values of prediction error as indicated by the asymmetry of the data around 0 residual and the bifurcation in the data on the positive end of the scale, corresponding to the two ends of the rating scale in Fig. 2(F). A similar pattern is apparent, although to a lesser extent, in Fig. 4(E), which represents errors of prediction for the task using a continuous scale without comparison stimuli [cf. Fig. 2(E), in which the *U* pattern is less marked than in Fig. 2(F)]. This pattern of residual error suggests that some of the remaining variance in levels of rater agreement may be due to an interaction between the mean rating and the presence/absence of comparison stimuli in determining the likelihood of listener agreement. To test this hypothesis, we repeated the ANCOVA analysis while omitting the two tasks without comparison stimuli. R^2 increased 4.1% to 88.3% variance accounted for. The effect of the covariate remained statistically significant, but the variance it accounted for decreased from 4.3% to 2.0%.

C. Patterns of agreement across voices

The above results indicate that the context in which a quality judgment is made (represented in this study by the presence and type of comparison stimuli) is the primary determinant of listener agreement levels. This result in turn suggested that in conditions where the comparison stimuli do not match the target voice, the task of assessing quality might be easier in some cases than others (and listeners might agree better in those cases) because of differences in the extent to which the generic comparison stimuli and test stimuli resembled each other, just by chance. To test this hypothesis, we measured the difference between the test stimuli and generic comparison stimuli with respect to the center frequencies of the first three formants, the difference in amplitude between the first and second harmonics (H1–H2), mean F0, and F0 sd, and used these differences to predict the probability of exact listener agreement in the experiments using generic comparison stimuli (Experiments 3 and 4). For the continuous rating scale task, the likelihood of listener agreement increased significantly with similarity in H1–H2 values [$F(1,38)=15.75$, $p<0.05$; $r=0.54$]. For the six-point scale task, the likelihood of exact listener agreement increased significantly with similarity in H1–H2 and the center frequency of F3 [$F(2,37)=5.50$, $p<0.05$; multiple $R=0.48$].

IV. DISCUSSION

This study tested the hypothesis that four factors (instability of memory standards for levels of a quality, ability to isolate single dimensions in a complex context, scale resolution, and the magnitude of the attribute being measured) govern the extent to which listeners do or do not agree in voice quality assessments. These four factors accounted for 84.2% of the observed variance in listener agreement levels, with all four factors contributing significantly to this value. Examination of residual errors of prediction (Fig. 4) did not show any other systematic tendencies, suggesting that no other significant factors contribute to predicting agreement levels. This means that a testing protocol that controls the present small set of factors should in fact provide the optimal tool for measuring voice quality, consistent with the theoretical framework described in the introduction. In fact, listeners agreed exactly 96% of the time in the task using custom comparison stimuli and a continuous scale, consistent with this prediction. Although the present study only examined ratings of breathiness and the NSR, we note that breathiness is integrally related to roughness and to overall hoarseness (Kreiman *et al.*, 1994), suggesting that the present results may generalize to ratings of other voice qualities and to the study of voice quality in general.

The numerical model presented in Table V quantifies the contributions of the different experimental factors and their interactions to facilitating listener agreement. Although the precise values associated with each term will probably vary somewhat with the attribute being measured and with the test voices under study, these coefficients can be interpreted as a model of the voice quality perception process. The most important factor in this model is the relative ease of isolating the quality under consideration in the different tasks, corresponding experimentally to provision of custom versus generic comparison stimuli. When comparison stimuli matched the test voices, the likelihood that listeners would agree exactly in their assessment of voice quality more than doubled compared to an unmatched context. When no external comparison stimuli are presented (as in traditional verbal rating scale tasks), listeners must rely on unstable internal standards for vocal qualities and rater agreement levels suffer (as quantified by negative values in Table V for tasks without comparison stimuli); but unstable internal standards are apparently better than mismatched external standards (contrary to our initial hypothesis). At best, providing comparison stimuli that did not match the target voices provided no advantage, and at worst including such stimuli in the assessment protocol actually reduced listener agreement levels relative to all other conditions. These results indicate that listeners can assess a given attribute of a voice pattern most reliably *in the context of the pattern itself*. A matching context facilitates this process; an internally generated standard overall has a negative impact; but providing an unmatched context confuses listeners and is a handicap, not a help, in collecting reliable measurements of voice quality.

Scale resolution also contributes significantly to measurement error: Listeners agreed significantly better on continuous scales than on six-point scales. Further, interactions between scale resolution and comparison stimulus condition explain the fact that agreement levels are similar for Experiments 4 (generic comparison stimuli, six-point scale) and 6 (no comparison stimuli, six-point scale) despite the large negative impact of generic comparison stimuli on agreement levels. As described in Sec. III, response biases were observed when six-point scales were combined with comparison stimuli: Listeners as a group preferred comparison stimuli with too little noise to those with too much noise, with agreement increasing as a result of this consistent response strategy. Simple examination of agreement rates out

of the context of the model in Table V would suggest that test protocols providing generic comparison stimuli are superior to protocols without comparison stimuli, but this impression is a result of the effects of this bias-derived interaction effect, and not of the comparison stimulus condition alone. This observation highlights the difficulty of understanding when and why listeners agree or disagree outside the context of some model of the voice perception process.

The approach taken in this article to the study of listener variability in quality assessment tasks stands in contrast to a recent proposal by Shrivastav and colleagues (Shrivastav *et al.*, 2005), who partitioned rating variance into two components: random errors and systematic biases (e.g., the tendency of one rater to rate all stimuli higher than another rater). They argued that random error can be eliminated post hoc by averaging together many ratings of a given voice from a single rater, after which systematic biases can be corrected by standardizing each listener's mean rating for the voice. In light of the present results, theoretical difficulties are apparent with this approach in addition to the practical limitations noted by the authors. The present results suggest that only a very small portion of rating variance is in fact random (exact agreement equaled 96% in Experiment 1, in which all hypothetical sources of variability were controlled), so that the additional advantage accrued by averaging a large number of ratings is small. Second, normalization of averaged ratings requires the assumption that errors due to response biases remain stable in nature and extent across some temporal testing window, so that there is actually some consistent effect for the normalization to correct. If these biases vary across the window of normalization, then the effect is not consistent, and normalization is inappropriate. Evidence suggests that this assumption is not always correct. Gerratt *et al.* (1993) showed that expert listeners' ratings can "drift" in predictable ways within a single rating session as listeners' unstable internal standards for levels of a quality adapt to the current listening context, so that a voice that sounded moderately rough on first hearing sounds severely rough a few minutes later in the context of many mildly deviant voices. In this case, ratings from one part of a listening session are not directly comparable to those from another part of the session.

The present results also shed light on the variable results produced by previous studies using constant comparison stimuli as perceptual anchors in protocols for assessing the breathiness and/or roughness of pathological voice stimuli (Chan and Yiu, 2002; Gerratt *et al.*, 1993). Such "anchored tasks" are attractive in practical terms, because they are easy to use and therefore potentially applicable in the clinic, but studies have found both increases (Gerratt *et al.*, 1993) and decreases (Chan and Yiu, 2002) in listener reliability relative to traditional unanchored voice quality ratings. In the context of the proposed model of the sources of variability in quality assessments, differences in study design account for these variable findings. Gerratt *et al.* (1993) used test and anchor stimuli drawn from the same synthetic continuum (a custom comparison stimulus task), and found that although agreement among raters improved overall (presumably due to the use of custom comparison stimuli), agreement among listeners decreased the more the target acoustic parameter in the test stimuli differed from that same parameter in the anchor stimuli, an apparent result of the poor resolution provided by their five-point scale. In contrast, listeners in Chan and Yiu (2002) referred to two generic comparison stimuli while performing a visual analog rating scale task. In that study, rating variability for untrained listeners remained unchanged or even increased relative to ratings made without comparison stimuli. This finding is consistent with the present finding that a generic anchor stimulus provides no benefit, presumably because listeners cannot reliably separate the quality being assessed from the overall voice pattern.

Finally, these results suggest that listener perception of voice quality involves a specific combination of holistic, gestalt-like pattern processing and featural analysis. To the extent that listeners rely on individual acoustic dimensions when making quality judgments, their attention to a dimension appears to depend on access to the context of the entire voice pattern. In other words, the features are part of the pattern, and have little or no perceptual importance outside of that pattern. This result is consistent with descriptions of the cognitive and neuropsychological processes underlying recognition of speaker identity from voice. In particular, behavioral data from patients with brain lesions and from control subjects indicate that the importance of individual acoustic cues to speaker identity in voice recognition tasks can be evaluated only in the context of the overall voice pattern (e.g., Van Lancker *et al.*, 1985; Van Lancker and Kreiman, 1987; Neuner and Schweinberger, 2000; cf. neuroimaging results in Belin *et al.*, 2000; Belin *et al.*, 2002; Levy *et al.*, 2001; Warren *et al.*, 2006; see Kreiman and Sidtis, 2008, for extended review). Voice recognition tasks involve assessing voice quality, so this apparent convergence is not surprising, and may point the way for development of a common theoretical framework for understanding the principles underlying many different aspects of voice perception. Studies of voice quality are nearly always specific to some particular discipline or research question (speaker recognition from voice; clinical assessment of voice quality; detection of lying from voice; assessment of the speaker's emotional state; and so on), and no comprehensive model exists to explain how listeners exploit different aspects of the voice signal when performing these various tasks. However, by considering the commonalities between studies of the different uses listeners make of voice quality, it may be possible to develop explanations that unify seemingly disparate areas of study.

ACKNOWLEDGMENTS

Diana Sidtis provided helpful comments on a preliminary draft, and comments by three anonymous reviewers resulted in significant improvements to this paper. Analysis and synthesis programs were written by Norma Antoñanzas-Barroso, with additional programming support from Brian Gabelman, Diane Budzik, and Ahror Rahmedov. All software used in this research is available as open source shareware at http://www.surgery.medsch.ucla.edu/glottalaffairs/. Research in the UCLA Voice Laboratory is supported by NIH/NIDCD Grant No. DC01797.

APPENDIX: COEFFICIENTS OF THE ANCOVA MODELS

$$A1 = \begin{cases} 1 & \text{if the task uses a six-point rating scale} \\ -1 & \text{when the rating scale is continuous,} \end{cases}$$

$$C1 = \begin{cases} 1 & \text{when the task uses generic comparison stimuli} \\ 0 & \text{when the task uses custom comparison stimuli} \\ -1 & \text{when the task does not include comparison stimuli,} \end{cases}$$

$$C2 = \begin{cases} 0 & \text{if the task uses generic comparison stimuli} \\ 1 & \text{if the task uses custom comparison stimuli} \\ -1 & \text{when the task does not include comparison stimuli,} \end{cases}$$

M = mean rating across listeners for that voice.

When all six experimental tasks are included in the analysis, the probability of exact listener agreement is predicted by

$$\text{Predicted agreement} = 0.3644 + (-0.0963 \times A1) + (-0.1608 \times C1) + (0.2810 \times C2) + (0.0535 \times A1 \times C1) + (-0.0268 \times A1 \times C2) + (0.3149 \times M)$$

Belin, P., Zatorre, R. J., and Ahad, P. (**2002**). "Human temporal-lobe response to vocal sounds," Brain Res. Cognit. Brain Res. **13**, 17–26.

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., and Pike, B. (**2000**). "Voice-selective areas in human auditory cortex," Nature (London) **403**, 309–312.

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.

Chan, K. M. K., and Yiu, E. M-L. (**2002**). "The effect of anchors and training on the reliability of perceptual voice evaluation," J. Speech Lang. Hear. Res. **45**, 111–126.

Cullinan, W. L., Prather, E. M., and Williams, D. E. (**1963**). "Comparison of procedures for scaling severity of stuttering," J. Speech Hear. Res. **6**, 187–194.

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.

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

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

Gerratt, B. R., and Kreiman, J. (**2007**). "Information conveyed by voices," paper presented at the 153rd Meeting of the Acoustical Society of America, June.

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.

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

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

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

Klatt, D. H., and Klatt, L. C. (**1990**). "Analysis, synthesis, and perception of voice quality variations among female and male talkers," J. Acoust. Soc. Am. **87**, 820–857.

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. (**2005**). "Perception of aperiodicity in pathological voice," J. Acoust. Soc. Am. **117**, 2201–2211.

Kreiman, J., Gerratt, B. R., and Antoñanzas-Barroso, N. (**2006**). "Analysis and synthesis of pathological voice quality," Technical report, available at http://www.surgery.medsch.ucla.edu/glottalaffairs/files/GASoftwareManual2006.pdf (last viewed July 11, 2007).

Kreiman, J., Gerratt, B. R., and Antoñanzas-Barroso, N. (**2007**). "Measures of the glottal source spectrum," J. Speech Lang. Hear. Res. **50**, 595–610.

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. B., 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.

Kreiman, J., and Sidtis, D. (**2008**). *Voices and Listeners* (Blackwell Publishing, Oxford).

Levy, D. A., Granot, R., and Bentin, S. (**2001**). "Processing specificity for human voice stimuli: Electrophysiological evidence," NeuroReport **12**, 2653–2657.

Ludlow, C. L. (**1981**). "Research needs for the assessment of phonatory function," ASHA Reports **11**, 3–8.

Macmillan, N. A., and Creelman, C. D. (**2005**). *Detection Theory: A User's Guide*,second edition (Cambridge University Press, Cambridge, UK).

Neuner, F., and Schweinberger, S. R. (**2000**). "Neuropsychological impairments in the recognition of faces, voices, and personal names," Brain and Cognition **44**, 342–366.

Shrivastav, R., and Sapienza, C. M. (**2003**). "Objective measures of breathy voice quality obtained using an auditory model," J. Acoust. Soc. Am. **114**, 2217–2224.

Shrivastav, R., Sapienza, C., and Nandur, V. (**2005**). "Application of psychometric theory to the measurement of voice quality using rating scales," J. Speech Lang. Hear. Res. **48**, 323–335.

Shrivastav, R., and Sapienza, C., (**2006**). "Some difference limens for the perception of breathiness," J. Acoust. Soc. Am. **120**, 416–423.

Titze, I. R. (**1994**). "Toward standards in acoustic analysis of voice," J. Voice **8**, 1–7.

Van Lancker, D., and Kreiman, J. (**1987**). "Voice discrimination and recognition are separate abilities," Neuropsychologia **25**, 829–834.

Van Lancker, D., Kreiman, J., and Emmorey, K. (**1985**). "Familiar voice recognition: Patterns and parameters. Part I: Recognition of backward voices," J. Phonetics **13**, 19–38.

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, Utrecht, pp. 12–15.

Warren, J. D., Scott, S. K., Price, C. J., and Griffiths, T. D. (**2006**). "Human brain mechanisms for the early analysis of voices," Neuroimage **31**, 1389–1397.

Webb, A. L., Carding, P. N., Deary, I. J., MacKenzie, K., Steen, N., and Wilson, J. A. (**2003**). "The reliability of three perceptual evaluation scales for dysphonia," Eur. Arch. Otorhinolaryngol. **261**, 429–434.

Yiu, E., and Ng, C. Y. (**2004**). "Equal appearing interval and visual analogue scaling of perceptual roughness and breathiness," Clin. Linguist. Phonetics **18**, 211–229.