Title
Validating a psychoacoustic model of voice quality.

Journal
The Journal of the Acoustical Society of America, 149(1)

ISSN
0001-4966

Authors
Kreiman, Jody
Lee, Yoonjeong
Garellek, Marc
et al.

Publication Date
2021

DOI
10.1121/10.0003331

Peer reviewed

JASA ARTICLE

Validating a psychoacoustic model of voice quality

Jody Kreiman,[1,a)] Yoonjeong Lee,[1] Marc Garellek,[2] Robin Samlan,[3] and Bruce R. Gerratt[4]

[1]*Departments of Head and Neck Surgery and Linguistics, University of California–Los Angeles, Los Angeles, California 90095-1794, USA*

[2]*Department of Linguistics, University of California–San Diego, San Diego, California 92093-0108, USA*

[3]*Department of Speech, Language, and Hearing Sciences, University of Arizona, Tucson, Arizona 85721, USA*

[4]*Department of Head and Neck Surgery, University of California–Los Angeles School of Medicine, Los Angeles, California 90095-1794, USA*

ABSTRACT:

No agreed-upon method currently exists for objective measurement of perceived voice quality. This paper describes validation of a psychoacoustic model designed to fill this gap. This model includes parameters to characterize the harmonic and inharmonic voice sources, vocal tract transfer function, fundamental frequency, and amplitude of the voice, which together serve to completely quantify the integral sound of a target voice sample. In experiment 1, 200 voices with and without diagnosed vocal pathology were fit with the model using analysis-by-synthesis. The resulting synthetic voice samples were not distinguishable from the original voice tokens, suggesting that the model has all the parameters it needs to fully quantify voice quality. In experiment 2 parameters that model the harmonic voice source were removed one by one, and the voice tokens were re-synthesized with the reduced model. In every case the lower-dimensional models provided worse perceptual matches to the quality of the natural tokens than did the original set, indicating that the psychoacoustic model cannot be reduced in dimensionality without loss of fit to the data. Results confirm that this model can be validly applied to quantify voice quality in clinical and research applications. https://doi.org/10.1121/10.0003331

(Received 23 June 2020; revised 7 December 2020; accepted 16 December 2020; published online 21 January 2021)

[Editor: Benjamin V. Tucker] Pages: 457–465

I. INTRODUCTION

At present, no agreed-upon method exists for objective measurement of perceived voice quality. As traditionally defined, voice quality is a psychoacoustic attribute—the perceptual response to all the acoustic attributes of a voice signal [ANSI (1960); see also Sundberg (1987) and Kreiman and Sidtis (2011)]. It follows that modeling or measuring voice quality entails identifying a set of acoustic attributes that are both necessary and sufficient to specify voice quality perception—a psychoacoustic level of description.

This is not the approach taken by traditional quality assessment protocols like the Consensus Auditory-Perceptual Evaluation of Voice (Kempster *et al.*, 2009) or the GRBAS protocol (Isshiki *et al.*, 1969), which partition voice quality into separate perceptual dimensions. In addition to scales like breathiness and roughness, these protocols typically include a scale for "grade" or overall severity of disorder, yet it is not clear how the individual quality scales relate to scaled severity or to the overall voice pattern. These protocols do not pretend to measure quality as a whole; and to our knowledge neither the necessity of individual scales nor the sufficiency of the composite protocols as models of overall quality has been established. As a result, two voices with identical profiles of ratings across scales can and do differ substantially in perceived overall quality. For this reason, it is impossible *a priori* for rating scales to provide information about how a listener actually perceives an overall voice pattern, one primary purpose of voice quality measurement. As a further limitation, listeners have difficulty focusing their attention on individual features like breathiness or roughness within complex acoustic patterns in voice, an inability that is the primary source of often-documented rating unreliability in traditional voice quality assessment protocols (Kreiman *et al.*, 2007). This further limits the effectiveness of scalar ratings of individual qualities as measures of the sound of a voice.

To address these issues, we have recently proposed an alternate model that treats quality as perceptually integral and models it as the set of acoustic parameters that allow listeners to determine that two signals are the same or different (Kreiman *et al.*, 2014). These parameters (Table I) were derived from a series of acoustic and psychoacoustic studies [Kreiman *et al.* (2007), Garellek *et al.* (2016), and Signorello *et al.* (2016)], and were selected because they account for most of the acoustic variability across voices. The assumption is that those parameters that vary most will be the most perceptually salient. This last point has not been formally examined, and the model as a whole thus remains to be validated.

Model validation requires demonstrating two things: That the model includes all the parameters needed to quantify a very wide range of voice qualities (i.e., the parameter set is sufficient), and that all included parameters are

[a)]Electronic mail: jkreiman@ucla.edu, ORCID: 0000-0002-5360-1729.

TABLE I. The parameters included in the psychoacoustic model of voice quality.

Model component	Parameters
Harmonic voice source	H1-H2, H2-H4, H4-2kHz, 2 kHz-5kHz
Inharmonic voice source	Spectral slope in four ranges (0–961 Hz, 961–2307 Hz, 2307–3653 Hz, 3653 Hz–5 kHz); HNR mean
Pitch	F0 mean; F0 contour
Loudness	Amplitude mean; amplitude contour
Vocal tract	Formants 1-11; bandwidths 1-11; spectral zeros 1-3; zero bandwidths 1-3

actually necessary. This paper presents two experiments addressing these points. Experiment 1 examines the sufficiency of the model—the range of phenomena for which it can account satisfactorily—by using analysis-by-synthesis to fit the model to a very wide range of naturally occurring voice qualities. Experiment 2 addresses the necessity of the parameters modeling the harmonic voice source (Table I) by eliminating them one by one and comparing the resulting synthesized voices to natural target voices.

II. EXPERIMENT 1

The goal of this experiment was to assess the limits of what the proposed psychoacoustic quality model can account for. Two hypotheses were tested: (1) model parameters will be sufficient to recreate the perceived quality of all normal and most pathological voices; (2) failures to adequately model quality will increase with increasing severity of perceived vocal pathology.

A. Method

1. Voice samples

One hundred voice samples (50 male, 50 female) were drawn from a database of recordings of speakers who had a voice disorder diagnosed by an otolaryngologist. Voices were unselected with respect to diagnoses[1] and ranged from extremely mild to very severe vocal pathology, as initially judged by the first author and confirmed via pretest (described next). An additional 100 voices (50 male, 50 female) were drawn from the UCLA Speaker Variability Database (Keating *et al.*, 2019), which includes multiple voice samples from over 200 male and female UCLA undergraduate students, none of whom reported a history of voice or speech complaints. All speakers sustained the vowel /a/ at comfortable pitch and loudness levels, and all were recorded with a Brüel and Kjær 1/2 in. microphone. Samples were directly digitized at 20 kHz (clinical samples) or 22 kHz (Speaker Variability Database samples) sampling rates. A relatively steady-state 1-s portion was selected from the middle of each utterance (so that onsets and offsets were eliminated). Samples were then downsampled to 10 kHz prior to analysis and testing.

2. Listening pretest

The following pretest was undertaken to confirm that the sample of voices included a wide range of severities of vocal pathology, and to provide data for testing the second hypothesis. All experimental procedures described in this paper were approved by the UCLA Institutional Review Board.

Pretest methods are fully described in Kreiman *et al.* (2020). Briefly, listeners judged the extent to which each natural voice sample was or was not normal using a visual sort and rate task (Granqvist, 2003). Male and female voices were judged separately, but samples from normal and clinical speakers were combined. The following procedure was used to control for context effects on perceived severity of dysphonia. Each listener heard 180 stimuli, either all male or all female, divided into 9 trials of 20 voices each. Five different sets of 180 stimuli were created for the male and female voices (10 sets total), such that across the entire experiment every voice appeared at least once in a trial with every other same-sex voice; 80 stimuli in each 180 voice set were repeated twice, and twenty appeared once only. No voices were repeated within a single 20 voice trial. Ten UCLA students and staff heard each of the ten 180-voice sets, for a total of 100 listeners (50 each for male and female voices). Listeners ranged in age from 18 to 68 years (mean age = 22.5 years; sd = 10.13 years). All listeners reported normal hearing. Students received course credit in return for their participation.

The experiment took place in a double-walled sound suite. Subjects were tested individually and heard the stimuli over Etymotic insert earphones (model ER-1; Etymotic Research, Inc., Elk Grove Village, IL) at a comfortable constant listening level. In each trial, listeners were presented with a screen containing 20 randomly colored and shaped icons, each icon representing a single voice token randomly assigned to that icon. Listeners played each voice by clicking its icon, then dragged the icon to a line at the bottom of the screen to indicate (1) whether the voice sounded normal, in which case the icon was placed in a box on the right end of the line, and (2) if it did not sound normal, how close to normal it sounded. The most abnormal-sounding voices were to be placed towards the left end of the line; those that approached normal were placed near the box. Voices judged equally dysphonic were to be stacked on the line so that they were the same distance from the ends of the line. Listeners were told that they could place as many or as few icons as desired in the box. They were encouraged to play the voices as often as required, in any order, until they were satisfied with their sort, after which testing advanced to the next trial. The experiment was self-paced and listeners could take breaks as needed. They were not told how many total speakers were included in the experiment. Testing lasted less than 1 h.

Icons placed in the box were assigned a rating of 1000; those at the left end of the line were scored 0, with scores for other icons interpolated between these values. Ratings

https://doi.org/10.1121/10.0003331

were averaged across listeners for use in calibrating results of the validation study. [For more detailed analyses, see Kreiman *et al.* (2020).] Rating distributions were skewed towards the right, consistent with the inclusion of samples from equal numbers of normal and pathologic speakers [Figs. 1(a) and 1(b)]. Across voices, mean ratings ranged from 122 to 917.9, where 0 meant maximally dysphonic and 1000 meant normal. No significant differences were observed between male and female speakers in mean ratings [$F(1, 198) = 0.15$, $p > 0.05$, $r^2 = 0$]. Listeners were quite self-consistent in their judgments (mean test-retest agreement = 75.8%; sd = 9.22%), but showed considerable between-listener variability [mean Spearman's rho for pairs of listeners = 0.27; sd = 0.11; see Kreiman *et al.* (2020), for discussion]. However, given the large number of ratings (n = 90) used to generate the mean values used here, and given that mean values spanned nearly the entire 1000 point scale, we conclude that the sample of voices was sufficiently large and varied to provide a fair test of the adequacy of the psychoacoustic model.

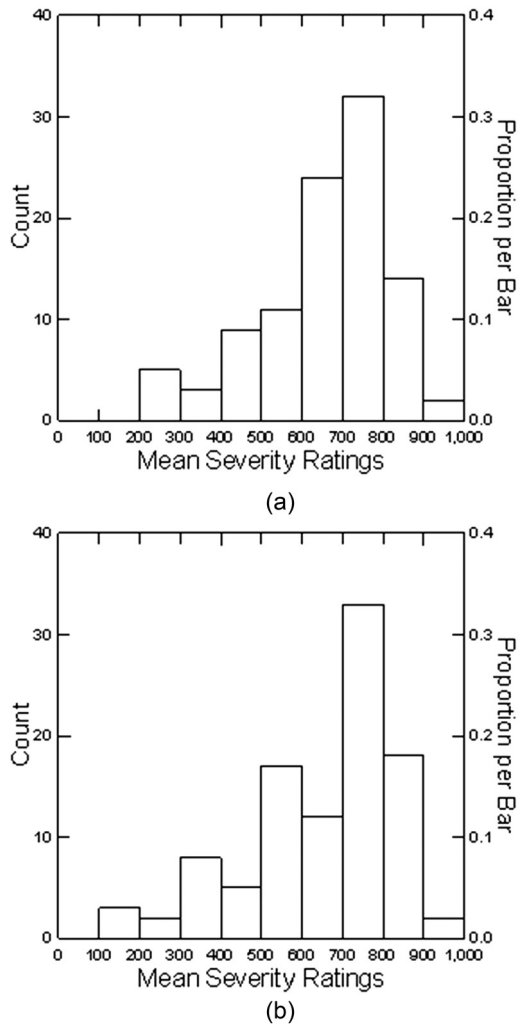

FIG. 1. Distribution of severity ratings across listeners. Larger values on the *x* axis represent more normal-sounding voices. (a) Female speakers. (b) Male speakers.

3. Synthesis procedures

Synthetic copies were created for each of the 200 voice samples using the UCLA voice synthesizer, which implements the psychoacoustic model of voice. All synthesis was completed by the first author. Methods are described in detail in Kreiman *et al.* (2016) [see also Kreiman *et al.* (2010)]. Briefly, voice samples were inverse filtered using the method described by Javkin *et al.* (1987). Harmonic source spectra were calculated from the resulting source pulses and then smoothed by fitting them with the model of the harmonic source spectrum (Table I, row 1), which models overall spectral shape in four pieces (H1 to H2; H2 to H4; H4 to the harmonic nearest 2 kHz; and the harmonic nearest 2 kHz to the harmonic nearest 5 kHz) but eliminates differences in amplitude between adjacent harmonics [Figs. 2(a) and 2(b)]. The inharmonic (noise) source spectrum was estimated through application of a cepstral-domain comb filter (a "lifter") like that described by de Krom (1993) [see also Qi and Hillman (1997)]. This spectrum was smoothed with a similar four-piece approximation [Figs. 2(c) and 2(d)], with segments spanning 0–961 Hz, 961–2307 Hz, 2307–3653 Hz, and 3653 Hz–5 kHz.

Fundamental frequency (F0) and amplitude contours were calculated from measurements of the original voice samples, and source pulses with frequencies and amplitudes dictated by these contours were calculated, then concatenated. A 100 tap FIR filter was synthesized for the noise spectrum, and a spectrally shaped noise time series was created by passing white noise through this filter. The source pulse train was added to this noise time series to create a complete glottal source time series.

The vocal tract was modeled by importing formant frequencies and bandwidths from the inverse filtering algorithm, and the complete synthesized source was filtered through this vocal tract model. The ratio of noise to harmonic energy was adjusted to approximate the value calculated from the original voice sample, resulting in a preliminary version of the synthetic voice. Finally, all parameters were adjusted to provide the best possible perceptual match (in the opinion of the first author) to the original voice sample. Although this procedure admits the possibility of using vocal tract parameters to compensate for insufficiencies in source parameterization, we note that above H4, formant changes have a very local effect relative to the wide frequency range of the source model segments, and thus are expected to have only a small effect on the higher-frequency harmonic slopes and their contribution to voice quality. We were less inclined to make formant changes that would affect frequencies below H4, because H1-H2 and H2-H4 have such narrow frequency bands that a change in formant frequency or bandwidth at this low frequency range would result in very large changes in quality (because they would affect both H1-H2 and H2-H4).

A dilemma arose during synthesis of tokens with changes in quality over the course of the token. The voice synthesizer was designed to model steady state phonation,

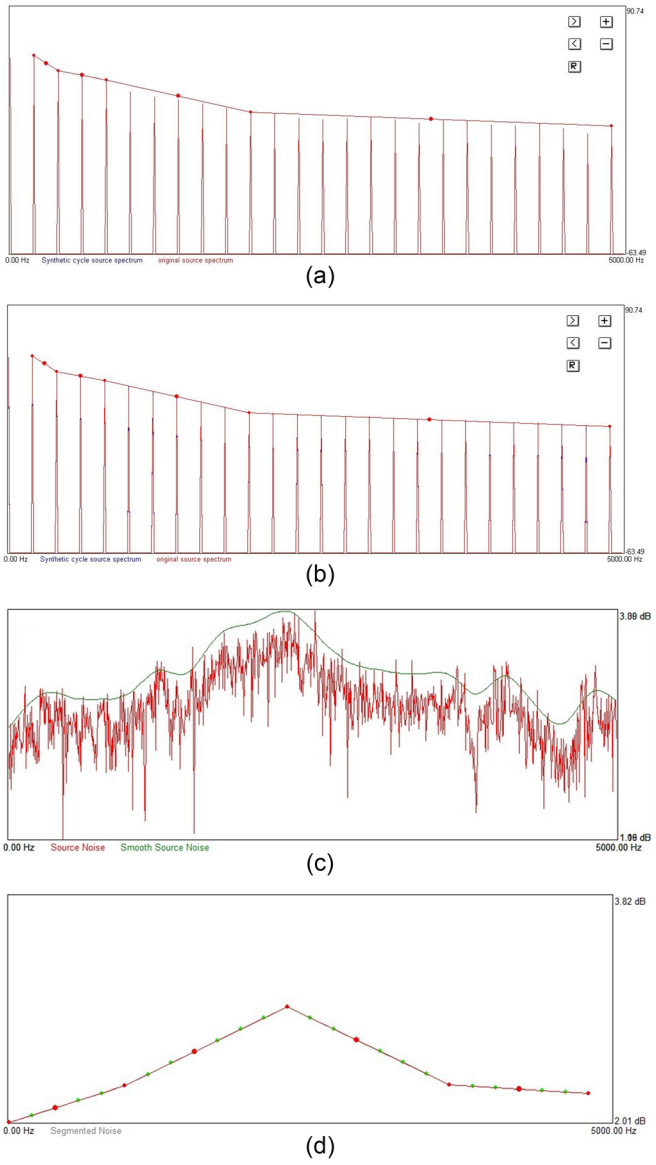

FIG. 2. (Color online) Parameterization of the harmonic and inharmonic voice sources. The x axis in these plots represents frequency in Hz; the y axis represents amplitude in dB. (a) The harmonic source spectrum before fitting the four-piece model. (b) The harmonic source spectrum after model fitting. (c) The inharmonic source spectrum. (d) The four-piece filter used to smooth this spectrum.

but voices with vocal pathology are often unsteady in quality, as are a fair number of tokens from speakers without obvious vocal pathology. This sometimes made it difficult to match the voice token precisely, even when the synthetic sample was a very good match to the speaker's overall voice quality (a "token vs type" problem). Although temporal details of such variations are important for matching the exact token under study, their relevance to the measurement of overall quality is less obvious, because such details are particular to a given sample and do not necessarily generalize to the overall sound of the voice. In response to this conflict, details of the particular sample were matched as closely as possible during synthesis, but our primary efforts were directed at capturing the speaker's individual voice quality. Listeners were asked to judge the samples with respect to both the match between tokens and the extent to which speakers' individual voice qualities were matched, as described in Sec. IV.

4. Perceptual evaluation

Synthetic and natural stimuli were combined to create 400 voice pairs: 200 where the two samples were identical ("same" pairs, randomly chosen to include either two natural samples or two synthetic samples), and 200 pairs where one voice was a natural sample and one was its synthetic copy ("different" pairs). Four different randomizations of these pairs were created, and each was divided into 2 blocks of 200 trials, for a total of 8 blocks of stimuli. Each block was judged by a separate group of 5 listeners, so that across blocks each "same" pair was judged by 20 listeners and each "different" pair was judged by 20 listeners, although these were not necessarily the same listeners. Listeners were drawn from the UCLA student population, and ranged in age from 18 to 29 years (mean = 19.2 years; sd = 1.90 years). All reported normal hearing. They received course credit for their participation.

Listeners were seated in a double-walled sound booth and heard the stimuli over Etymotic ER-1 insert earphones. On each trial, they heard the two 1-s stimuli, separated by 250 ms. They were allowed to play each pair of voices once in each order (A/B and B/A), after which they were asked to judge whether the two samples were identical (sample matching task), and to provide their confidence in their rating on a 1–5 scale where 1 meant they were positive about their response and 5 meant it was a wild guess. To assess the extent of the match to the quality of the voice independent of the temporal details of the voice sample, listeners also judged whether the two samples represented the same talker (talker matching task), although not necessarily the same sample from that speaker, again making confidence ratings on a 1–5 scale. In this case, judgments required listeners to ignore details of the voice sample, and instead decide whether differences between the samples were consistent or not with expected within-speaker variability in voice quality.

For both the sample and talker matching tasks, same/different sample or talker responses were combined with confidence ratings to create a single 10-point scale ranging from 1 (positive voices are the same; confidence rating = 1 and response = "same"), through 5 (unsure voices are the same; confidence rating = 5 and response = "same") and 6 (unsure voices are different; confidence rating = 5, response = "different") to 10 (positive voices are different; confidence rating = 1 and response = "different"). SYSTAT software (Version 13.1; Systat Software, Inc., San Jose, CA) was then used to calculate d′ from these unfolded confidence ratings. d′ values increase with increasing discrimination performance; a d′ value of 2.10 corresponds to 75% probability of a correct response [MacMillan and Creelman (2005), p. 385], and was used as a criterion for interpreting these results.

https://doi.org/10.1121/10.0003331

B. Results

Across voices, no listener performed at or above criterion levels, whether discriminating between synthetic and natural tokens or making same/different talker judgments. When discriminating between synthetic and natural tokens of the voices, d′ averaged 0.81 (sd = 0.50, range = –0.14 to 1.86). When asked if tokens represented the same or different speakers, d′ averaged 0.42 (sd = 0.46, range = –0.43 to 1.34).

Across listeners, discrimination scores were below criterion levels for all female voices, for both tasks (same/different sample task: mean d′ = 0.54, range = –0.45 to 1.99, sd = 0.39; same/different talker task: mean d′ = 0.19, range = –2.57 to 1.17, sd = 0.54). Overall performance was quite poor for male voices as well (mean d′ = 0.55, range = –0.44 to 2.4, sd = 0.47), but synthetic and natural tokens for two of the 50 pathologic talkers were discriminable at above criterion levels (d′ = 2.28 and 2.4). Figure 3 shows how values of the cepstral peak prominence (CPP) (Hillenbrand *et al.*, 1994), a computationally robust variant of the noise-to-harmonics ratio, vary over time for these two natural voice samples and for their synthetic counterparts. CPP values were calculated using VOICESAUCE software (Shue *et al.*, 2011), with a Hamming window five pitch periods in length. As these figures show, noise levels for the natural stimuli [Fig. 3(a) and 3(b)] increased and decreased more over time than for the synthetic samples [Fig. 3(c) and 3(d)], an impression that was confirmed by careful listening. However, accuracy did not exceed criterion levels for the same/different talkers task for any male voices, including these two (d′ = 1.14 and 0.87, respectively) (mean d′ = 0.18, range = –0.87 to 1.22; sd = 0.4).

Finally, contrary to expectation, we observed little or no relationship between perceived severity of voice disorder and discriminability of the natural and synthetic tokens (same/different token task: $r = -0.28$, $p < 0.01$; same/different talker task: $r = -0.1$, n.s.). Although the first correlation is statistically reliable, the amount of variance accounted for is negligible ($r^2 = 0.078$).

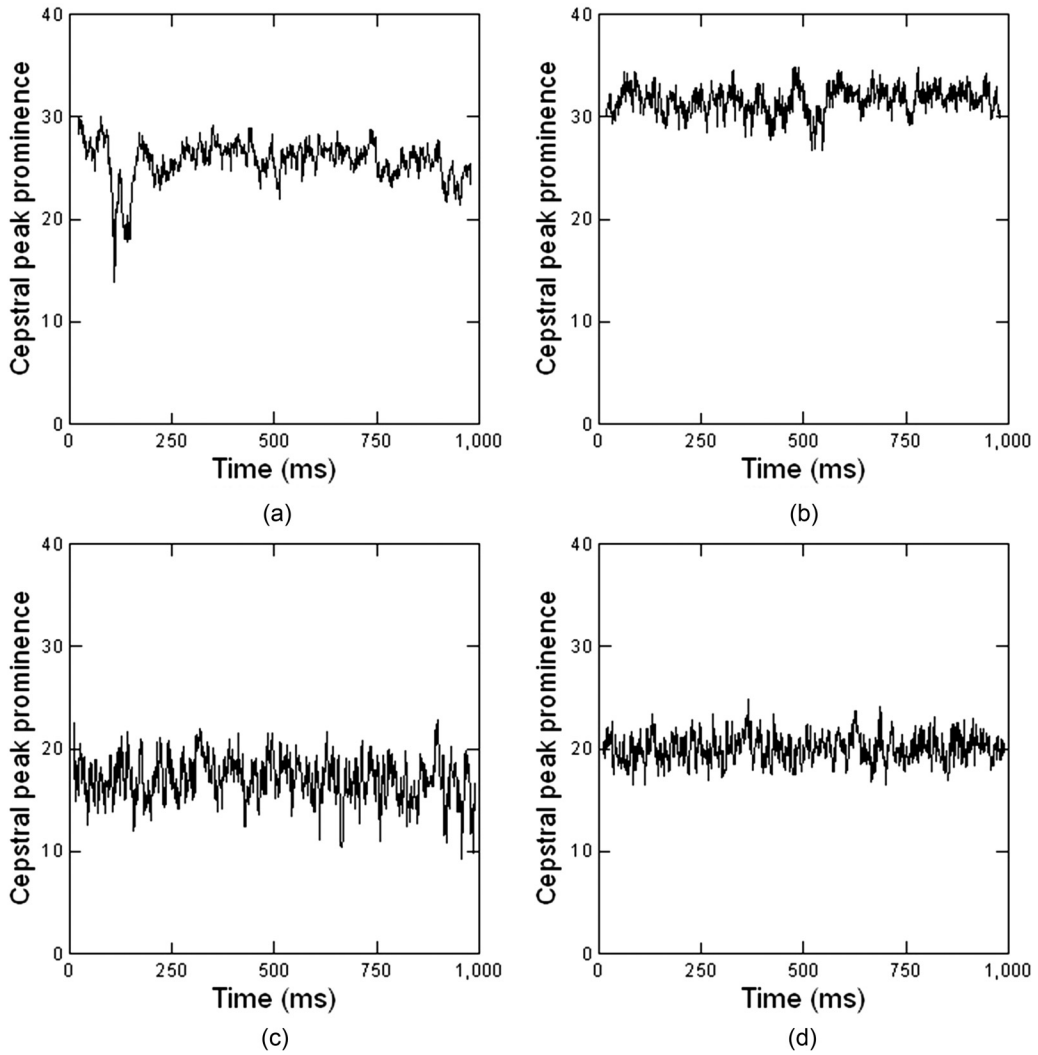

FIG. 3. Noise variability for two voice samples that were perceptually discriminable from their synthetic counterparts. (a) and (b) CPP values over time for the natural voice tokens. (c) and (d) CPP values over time for the synthetic stimuli corresponding to the natural voices in panels (a) and (b), respectively.

C. Interim discussion

It is of course impossible to prove that the psychoacoustic model of voice quality is adequate to model every possible voice, because of the obvious limits imposed by sampling. For this study, every effort was made to include a large range of vocal pathologies and severities of voice disorders, but even the large sample of voices studied here cannot capture the full range of possible qualities that exist or could exist, either within or between speakers. We also note that our sample contained a large number of normal voices, where "normal" was self-defined by the speakers. In a universe of talkers, the vast majority of voices are normal in this sense. Given these circumstances, the fact that only 2/200 synthetic tokens were reliably (but far from perfectly) discriminable from their natural counterparts is in our opinion strong evidence that the model adequately quantifies voice quality, particularly since those two tokens were judged to match the overall quality of the speaker, if not the exact temporal details of the specific voice sample. Further, for the current sample of voices, accuracy of the model did not decrease with increasing severity of dysphonia.

This leads to the second issue arising from these results. The two synthesis failures that occurred do not appear to be the result of model limitations, but rather were related to minor issues with token unsteadiness. This raises the question of what exactly we are attempting to model: the overall sound of a sample, or the precise details of its temporal variation. To our knowledge, this issue has not been addressed in studies evaluating previous protocols for quality assessment. The nature and extent of variability in voice in general are poorly understood [e.g., Lavan *et al.* (2019) and Lee *et al.* (2019)], as are the ways in which listeners cope with such variability when judging quality. A solution to this issue is beyond the scope of the present study; however, the psychoacoustic model proposed here may offer a tool for future work addressing this topic. We return to this in Sec. IV.

III. EXPERIMENT 2

Experiment 1 provided evidence that the psychoacoustic model has all the parameters it needs to model a wide range of voice qualities. This experiment addresses the remaining question about the model's validity: Are all the included parameters actually needed to model quality adequately?

In addressing this question, we assume that previous research has sufficiently established the perceptual importance of F0, formant frequencies and bandwidths, and sound intensity, so that their inclusion in the model need not be justified anew [see, e.g., Fastl and Zwicker (2007) and Hillenbrand (2019) for review]. Our own previous studies (Kreiman and Gerratt, 2005, 2012; Signorello *et al.*, 2016) have also established the importance of correctly modeling the inharmonic voice source. However, the necessity of all four parameters in the model of the harmonic source has not been previously established. These parameters were derived from acoustic analyses of a large number of voice sources, and were chosen so that they accounted for as much variance as possible in source spectral shape across different voices (Kreiman *et al.*, 2007). To test the hypothesis that perception of voice exploits acoustic variability (in other words, that listeners use the parameters that vary most across voices when they assess the quality of an individual voice), we created stimuli by dropping each piece in turn out of the harmonic source model, re-synthesizing the voices, and then assessing the effect of these changes on the match between synthetic and natural tokens. If acoustic variability predicts perception, then across voices, the four-piece source spectral model should provide a better match to the natural voice tokens than any of the three-piece models.

A. Methods

1. Stimuli

Twenty-four voices (12 male, 12 female) were selected from the voices used in experiment 1, such that one male and one female voice had a low, mid, or high value for each of the 4 spectral source parameters, based on the observed distribution of values for the entire set. Four versions of each stimulus voice were then created. The first was the token created with the four-parameter source spectral model via analysis-by-synthesis in experiment 1 [Fig. 4(a)]. In the second version, H1-H2 and H2-H4 were merged to create a single H1-H4 parameter [Fig. 4(b)]; in the third, H2-H4 was merged with H4–2 kHz to create an H2–2 kHz parameter [Fig. 4(c)]; and in the fourth, H4–2 kHz was merged with 2 kHz-5 kHz to create a single H4–5 kHz parameter [Fig. 4(d)]. As in experiment 1, differences in the amplitudes of individual harmonics within each range were eliminated.

In many cases, these changes to the harmonic source resulted in prominent changes in vowel quality, because the vocal tract models used in experiment 1 were optimized for a four-piece source model. For this reason, formant frequencies and bandwidths were re-adjusted to provide the best possible match to the natural target voice in the context of each of the 3 new harmonic sources, so that any mismatches in overall quality between the synthetic and natural tokens could be unambiguously attributed to differences among source models. Levels for the noise-to-harmonics ratio were also reset, to compensate for changes in the perceptual prominence of spectral noise as a result of changes to the harmonic source spectrum (Kreiman and Gerratt, 2012; Labuschagne and Ciocca, 2020). All other model parameters remained unchanged from their values in experiment 1.

2. Participants

Twenty listeners (12 female, 8 male) participated in this experiment. They ranged in age from 18 to 65 (mean = 31; sd = 14.3). All reported normal hearing. They were compensated for their time.

https://doi.org/10.1121/10.0003331

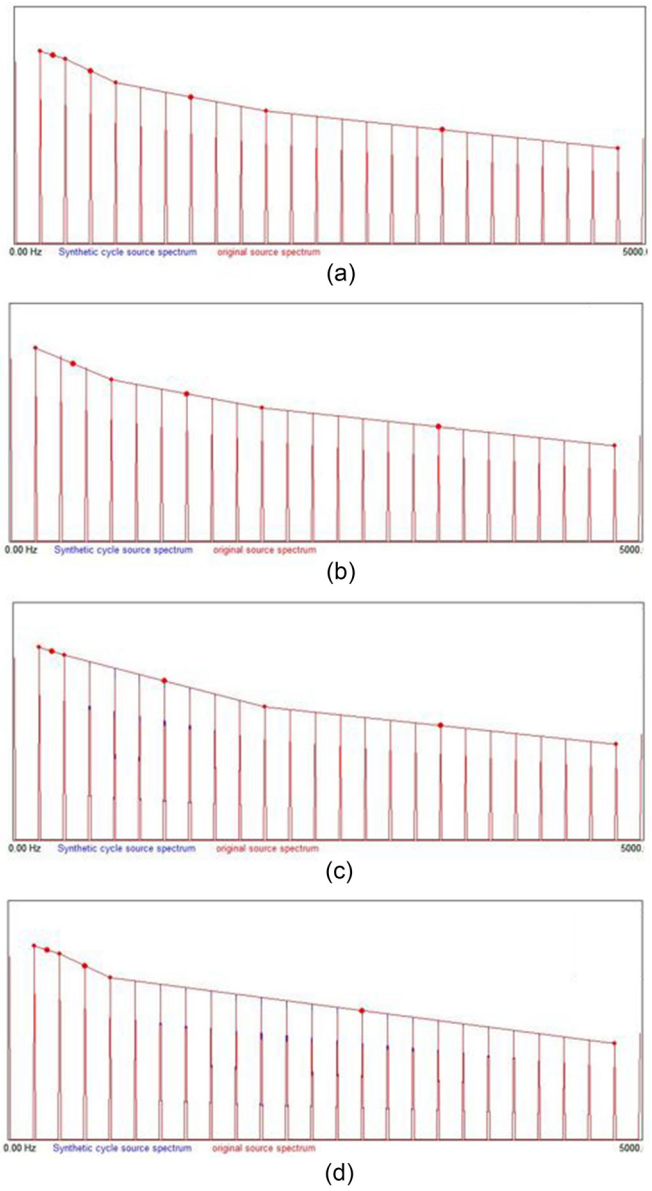

FIG. 4. (Color online) The three-piece source models. (a) A source fitted with the original four-piece model. (b) The same source fitted with a three-piece model using H1-H4. (c) The same source fitted with a three-piece model using H2-2 kHz. (d) The same source fitted with a three-piece model using H4-5 kHz.

3. Task

Listeners were seated in a double-walled sound booth and heard the stimuli over Etymotic ER-1 insert earphones. On each trial, they heard two 1-s stimuli, separated by 250 ms, and were asked to report whether the stimuli were the same or different along with their confidence in their response on a 1 (positive) to 5 (wild guess) scale. Stimuli were identical for half of the trials; for the other half, one stimulus was the natural voice token and one was one of the 4 synthetic versions of that voice (24 voices × 4 versions = 96 "different" trials, plus 96 "same" trials, for a total of 192 trials/listener). In all cases, playback was limited to 2 repetitions, once in each order (A/B and B/A). Testing lasted an average of about 40 min.

B. Results

As in experiment 1, "same" and "different" responses were combined with confidence ratings to create a single scale ranging from 1 (positive voices are the same) to 10 (positive voices are different). Data from all listeners were combined to calculate a single d′ value for each of the 96 natural/synthetic voice pairs. Results averaged across voices are given in Table II. Matched pair t-tests indicated that the four-piece source model provided a significantly better match (as measured by lower d′) to the natural voice tokens than did any of the three-piece models (four-piece model vs three-piece model with H1-H4: $t(23) = -3.69$, $p < 0.001$; four-piece model vs three-piece model with H2–2 kHz: $t(23) = -2.98$, $p < 0.007$; four-piece model vs three-piece model with H4–5 kHz: $t(23) = -3.70$, $p < 0.001$).

C. Discussion

Results were consistent with our hypothesis, in that the four-piece model of the harmonic voice source provided a better overall fit to quality than did any of the three-piece models. However, changes to different parts of the harmonic source spectrum had different effects on the quality of the synthetic stimuli. Changes to the detail with which the lowest harmonics were modeled (by merging H1-H2 and H2-H4 or H2-H4 and H4–2 kHz) significantly impacted vowel quality, which could be largely corrected by adjusting formant frequencies and bandwidths. Given that our stimuli were /a/ vowels, this is not surprising: The primary determinants of vowel quality, F1 and F2 (both generally between 700 and 1200 Hz for /a/) can change markedly in amplitude as the source spectral shape is modified in this range. The relationship between vowel quality and voice quality depends in theory on whether one views voice primarily from the perspective of production or perception. From a production point of view, researchers have long

TABLE II. d′ values for comparisons between natural stimuli and 3- vs 4-piece source models. Higher values represent better discrimination performance.

	Natural token vs. four piece model	Natural token vs. model with merged H1-H2 and H2-H4	Natural token vs. model with H2-2 kHz merged H2-H4 and H4-2kHz	Natural token vs. model with H4-5 kHz merged H4-2kHz and 2kHz-5kHz
Mean	1.48	1.92	1.95	2.25
SD	0.58	0.92	1.03	0.94
Minimum	0.40	0.71	0.58	0.77
Maximum	2.48	3.74	5.06	4.30

distinguished narrow from broad definitions of voice [e.g., Kreiman and Sidtis (2011)]: In a narrow view, voice comprises only those attributes directly related to the voice source (i.e., laryngeal activity), while in broader views, voice is nearly synonymous with speech and thus includes vocal tract resonances. Narrow definitions are uncommon in perceptual research, because listeners do not have separate access to laryngeal and resonance aspects of phonation (although attributes like "breathiness" or "roughness" are often assumed to be laryngeal in origin). By including formant frequencies and bandwidths as part of the psychoacoustic model of voice quality, we have implicitly adopted the view that effects of the vocal tract filter are part of perception of voice quality. However, the present results suggest that even from the standpoint of production strict separation of source and vocal tract functions is problematic: Speakers must adjust source and filter jointly if they are to simultaneously achieve both voice quality and vowel quality goals. This is inconsistent with the distinction between narrow and broad definitions of voice, and suggests that very narrow definitions of voice quality may be untenable.

Changes to the higher part of the harmonic source spectrum had less impact on vowel quality, but significantly affected the "breathy/turbulent" quality of the voice and overall brightness. Adjusting formant frequencies and/or bandwidths did not correct these quality changes, presumably because there are few formants relative to the number of harmonics in these larger frequency ranges. This implies that either (1) as we talk we make constant, small adjustments to the voice source to maintain a relatively constant voice quality as vowel quality changes in speech or (2) voice quality varies within technically perceptible ranges across utterances, but speakers and listeners are both focused on semantic meaning so they simply do not notice this. Data examining within-speaker variability in voice quality in connected speech are needed to untangle these issues. We note that the psychoacoustic model is an essential prerequisite to such studies, because it limits the number of acoustic features that need to be examined to a relatively small necessary and sufficient set.

IV. GENERAL DISCUSSION

Because models are summary descriptions of a universe of data, no model can account for every possible observation in its domain. With that said, the psychoacoustic model proposed here appears to account for a very large range of voice qualities in an economical manner, particularly when compared to Voice Profile Analysis (Laver, 1980; Laver *et al.*, 1981), to our knowledge the only perceptual protocol that purports to fully quantify voice quality—albeit in the production domain—which requires perceptual ratings of 36 parameters.[2] In addition to its relative simplicity, the proposed psychoacoustic protocol differs from scalar protocols in its approach to measuring voice quality, in that it quantifies the voice pattern as a whole, not as a set of individual attributes. Rating scale protocols focus on single attributes of voice quality like breathiness and roughness, and scores on individual scales are usually assumed to be meaningful out of the context of any other attributes the voice may have. In contrast, parameters of the psychoacoustic model (although statistically largely independent) are designed to quantify a complete integral pattern, and are not necessarily interpretable individually.

A focus on measuring overall voice quality could support clinical approaches that focus on treating the overall sound of a voice, rather than on individual dimensions like breathiness or roughness. Such approaches are intuitively appealing, but developing them requires linking the complete sound of a voice to the specific underlying pathology, an ambitious goal for the future. For the moment we note that linking voice production to perception is potentially an easier task in the context of a model that links acoustics to perception, facilitating further linkages back to the underlying vocal physiology.

One significant limitation of the psychoacoustic model is that at present it describes only steady-state phonation, except that F0 and amplitude variability are included in our model of the harmonic voice source (which calculates individual pulse periods and amplitudes based on tracks of the original sample). This limitation was imposed during model development for pragmatic reasons. As discussed above, studies of voice quality have not consistently distinguished the quality of a particular voice token from the overall sound of a speaker's voice, so that it is often unclear what exactly is being measured, and the theoretical status and proper quantification of variability in voice remain poorly understood. The relationship between within-sample versus within-speaker variability in quality should in principle derive from models of within-speaker variability in voice, but to our knowledge no such model exists at present. Informal observations suggest that variability in these and other model parameters is well quantified by coefficients of variation for the relevant parameters, but further research is needed to clarify these issues.

Finally, it is possible that another set of acoustic parameters exists that would describe voice quality equally well; and it is possible that more than one set of parameter values from the present set could result in equally good models of the target voice quality. Two factors minimize these concerns. First, the parameters included in the model were derived from extensive acoustic analyses of a very large number of test stimuli [including Kreiman *et al.* (2007), Garellek *et al.* (2016), and Signorello *et al.* (2016); summarized in Kreiman *et al.* (2014)]. As such, the parameter set provides a detailed acoustic model of the stimuli that accounts for much of the acoustic variability that distinguishes speakers. Second, this demonstration of the validity of the set of parameters allows us to measure parameters via automatic acoustic analysis [for example, with VOICESAUCE software; Shue *et al.* (2011) and Lee *et al.* (2019)] rather than requiring subjective estimation via analysis-by-synthesis. Concerns about parameter validity previously limited the

suitability of automatic acoustic analysis, but the existence of a valid psychoacoustic model eliminates these concerns; and use of automatic estimation procedures limits concerns about multiple solutions, because automatic procedures yield the same values each time they are applied.

In conclusion, although both theoretical and practical questions remain about how voice quality should be defined, this model of the relationship between acoustics and voice quality appears to validly quantify the sound of a wide range of voices, from normal to severely pathologic. Availability of such a measurement protocol may facilitate many future investigations of voice, including devising models relating voice production to voice perception. Only when we understand how physiological changes are related to changes in the sound of a voice, and vice versa, can we truly say we know why a voice sounds as it does.

ACKNOWLEDGMENTS

This research was supported by Grant No. DC01797 from the National Institute on Deafness and other Communication Disorders, and by Grant No. IIS-1704167 from the National Science Foundation. Voice synthesizer and testing software were written by Norma Antoñanzas, and are freely available on request to the first author, as is the UCLA Speaker Variability Database.

[1]Information about specific diagnoses is unavailable due to privacy regulations.

[2]Another possible exception, the GRBAS protocol (Isshiki *et al.*, 1969) was only designed to quantify the quality of hoarse voices, and thus cannot be said to measure overall voice quality in any general way.

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

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.

Fastl, H., and Zwicker, E. (**2007**). *Psychoacoustics: Facts and Models* (Springer, Berlin), pp. 111–148, 203–238.

Garellek, M., Samlan, R., Gerratt, B. R., and Kreiman, J. (**2016**). "Modeling the voice source in terms of spectral slopes," J. Acoust. Soc. Am. **139**, 1404–1410.

Granqvist, S. (**2003**). "The visual sort and rate method for perceptual evaluation in listening tests," Logoped. Phoniatr. Vocol. **28**, 109–116.

Hillenbrand, J. (**2019**). "The acoustics and perception of North American English vowels," in *The Routledge Handbook of Phonetics*, edited by W. F. Katz and P. F. Assmann (Routledge, London), pp. 219–263.

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

Isshiki, N., Okamura, H., Tanabe, M., and Morimoto, M. (**1969**). "Differential diagnosis of hoarseness," Folia Phoniatr. **21**, 9–19.

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

Keating, P., Kreiman, J., and Alwan, A. (**2019**). "A new speech database for within- and between-speaker variability," in *Proceedings of the 19th International Congress of Phonetic Sciences*, Melbourne, Australia, Australasian Speech Science and Technology Association Inc., Canberra, Australia, pp. 736–739.

Kempster, G. B., Gerratt, B. R., Verdolini Abbott, K., Barkmeier-Kraemer, J. M., and Hillman, R. E. (**2009**). "Consensus auditory-perceptual evaluation of voice: Development of a standardized clinical protocol," Am. J. Speech Lang. Pathol. **18**, 124–132.

Kreiman, J., Antoñanzas-Barroso, N., and Gerratt, B. R. (**2010**). "Integrated software for analysis and synthesis of voice quality," Behav. Res. Methods **42**, 1030–1041.

Kreiman, J., Auszmann, A., and Gerratt, B. R. (**2020**). "What does it mean for a voice to sound 'normal?,' " in *Voice Attractiveness: Studies on Sexy, Likable, and Charismatic Speakers*, edited by B. Weiss, J. Trouvain, M. Barkat-Defradas, and J. Ohala (Springer, Singapore), pp. 83–100.

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

Kreiman, J., and Gerratt, B. R. (**2012**). "Perceptual interactions of the harmonic source and noise in voice," J. Acoust. Soc. Am. **131**, 492–500.

Kreiman, J., Gerratt, B. R., and Antoñanzas-Barroso, N. (**2016**). "Analysis and synthesis of pathological voice quality," 2nd ed., available at http://headandnecksurgery.ucla.edu/glottalaffairs (Last viewed 12/31/2020).

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., Garellek, M., Samlan, R., and Zhang, Z. (**2014**). "Toward a unified theory of voice production and perception," Loq. Spanish J. Speech Sci. (published online).

Kreiman, J., Gerratt, B. R., and Ito, M. (**2007**). "When and why listeners disagree in voice quality assessment tasks," J. Acoust. Soc. Am. **122**, 2354–2364.

Kreiman, J., and Sidtis, D. (**2011**). *Foundations of Voice Studies: An Interdisciplinary Approach to Voice Production and Perception* (Wiley-Blackwell, Walden, MA), pp. 6–10.

Labuschagne, I. B., and Ciocca, V. (**2020**). "The effect of vocal tract parameters on aspiration noise discrimination," J. Acoust. Soc. Am. **147**, 1239–1249.

Lavan, N., Burston, L. F. K., and Garrido, L. (**2019**). "How many voices did you hear? Natural variability disrupts identity perception from unfamiliar voices," Br. J. Psychol. **110**, 576–593.

Laver, J. (**1980**). *The Phonetic Description of Voice Quality* (Cambridge University Press, Cambridge), pp. 157–165.

Laver, J., Wirz, S., Mackenzie, J., and Hiller, S. M. (**1981**). "A perceptual protocol for the analysis of vocal profiles," Edinburgh Univ. Dept. Linguistics Work Prog. **14**, 139–155.

Lee, Y., Keating, P., and Kreiman, J. (**2019**). "Acoustic voice variation within and between speakers," J. Acoust. Soc. Am. **146**, 1568–1579.

Macmillan, N. A., and Creelman, C. D. (**2005**). *Detection Theory: A User's Guide*, 2nd ed. (Lawrence Erlbaum Associates, Mahwah, NJ), p. 385.

Qi, Y., and Hillman, R. E. (**1997**). "Temporal and spectral estimations of harmonics-to-noise ratio in human voice signals," J. Acoust. Soc. Am. **102**, 537–543.

Shue, Y.-L., Keating, P., Vicenik, C., and Yu, K. (**2011**). "VoiceSauce: A program for voice analysis," In *Proceedings of the ICPhS XVII*, pp. 1846–1849.

Signorello, R., Rhee, N., Gerratt, B. R., and Kreiman, J. (**2016**). "Toward a psychoacoustic model of spectral noise in the voice source," presented at the *10th International Conference on Voice Physiology and Biomechanics (ICVPB)*, Vina del Mar, Chile.

Sundberg, J. (**1987**). *The Science of the Singing Voice* (Northern Illinois University Press, DeKalb, IL), pp. 1–5.