JSLHR

Research Article

Comparing Measures of Voice Quality From Sustained Phonation and Continuous Speech

Bruce R. Gerratt,[a] Jody Kreiman,[a] and Marc Garellek[b]

Purpose: The question of what type of utterance—a sustained vowel or continuous speech—is best for voice quality analysis has been extensively studied but with equivocal results. This study examines whether previously reported differences derive from the articulatory and prosodic factors occurring in continuous speech versus sustained phonation.
Method: Speakers with voice disorders sustained vowels and read sentences. Vowel samples were excerpted from the steadiest portion of each vowel in the sentences. In addition to sustained and excerpted vowels, a 3rd set of stimuli was created by shortening sustained vowel productions to match the duration of vowels excerpted from continuous speech. Acoustic measures were made on the stimuli, and listeners judged the severity of vocal quality deviation.
Results: Sustained vowels and those extracted from continuous speech contain essentially the same acoustic and perceptual information about vocal quality deviation.
Conclusions: Perceived and/or measured differences between continuous speech and sustained vowels derive largely from voice source variability across segmental and prosodic contexts and not from variations in vocal fold vibration in the quasisteady portion of the vowels. Approaches to voice quality assessment by using continuous speech samples average across utterances and may not adequately quantify the variability they are intended to assess.

The study of voice quality perception typically requires acoustic recordings of voice samples for analysis. However, the question of what type of voice sample is most appropriate for acoustic and/or perceptual analysis has been controversial, resulting in a number of studies examining the relative suitability of sustained vowels versus samples of continuous speech (e.g., Lederle, Barkmeier-Kraemer, & Finnegan, 2012; Maryn & Roy, 2012; Moon, Chung, Park, & Kim, 2012; Zraick, Wendel, & Smith-Olinde, 2005). Arguments in favor of measuring voice quality from sustained vowels contend they are relatively time invariant; free from influences of phonetic context and thereby unaffected by intonation, stress, or speaking rate; easy to elicit, produce, and analyze; more easily controlled; and less affected by the dialect of the speaker and/or listener than continuous speech is (Maryn, Corthals, Van Cauwenberge, Roy, & De Bodt, 2010a; Zraick et al., 2005). The relative absence of articulatory and prosodic influences may also help the listener focus more closely on aspects of quality related solely to the voice source (de Krom, 1994), reducing variability in listeners' perceptual responses.

Measures from continuous speech derive from broader conceptions of voice quality as nearly synonymous with speech, so that voice in this sense includes aspects of articulation (e.g., breathiness near /h/) and accent, unvoiced portions of utterances, sentential prosody (e.g., phrase-final creak, fundamental frequency declination), gestures related to linguistic voicing contrasts (e.g., breathiness due to aspiration at the release of voiceless aspirated stops and creak near /ʔ/), and so on. Arguments for assessing quality from continuous speech assert that it is more ecologically valid, because it better represents the dynamic attributes of voice that occur in regular speech (Maryn et al., 2010a; Parsa & Jamieson, 2001), such as vocal fluctuations that may present during voicing onset and termination and variations in amplitude and frequency (Awan, Roy, Jette, Meltzner, & Hillman, 2010). In this view, features such as voicing onsets, offsets, and contextual effects, the absence of which makes sustained phonation valuable, are instead considered essential to evaluation of voice in actual communicative use, while

[a]University of California, Los Angeles
[b]University of California, San Diego, La Jolla
Correspondence to Bruce R. Gerratt: bgerratt@ucla.edu

Editor: Rhea Paul
Associate Editor: Kate Bunton
Received September 2, 2015
Revision received December 10, 2015
Accepted March 24, 2016
DOI: 10.1044/2016_JSLHR-S-15-0307

Disclosure: *The authors have declared that no competing interests existed at the time of publication.*

steady-state vowels are viewed as relatively impoverished sources of information about vocal function (Maryn et al., 2010a). Further, speech elicited in specific phonetic contexts, such as those used in the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V; Kempster, Gerratt, Verdolini Abbott, Barkmeier-Kraemer, & Hillman, 2009), can reveal the existence and nature of some voice disorders (e.g., vocal tremor [Lederle et al., 2012] and adductory spasmodic dysphonia) that may occur more commonly under certain circumstances (at voicing onsets and/or offsets, for example, Awan et al., 2010; Roy, Gouse, Mauszycki, Merrill, & Smith, 2005). However, most acoustic studies of quality assessment from continuous speech use means and/or standard deviations for acoustic measures calculated across the entire sample of speech, and perceptual studies usually assess the overall extent of dysphonia (e.g., Awan, Roy, & Dromey, 2009; de Krom, 1994; Halberstam, 2004; Lederle et al., 2012; Lowell, Colton, Kelley, & Hahn, 2011; Maryn et al., 2010a; Maryn & Roy 2012; Moon et al., 2012; Parsa & Jamieson, 2001; Revis, Giovanni, Wuyts, & Triglia, 1999; Watts & Awan, 2011; Zraick et al., 2005; see Lowell, 2012, or Maryn, Roy, De Bodt, Van Cauwenberge, & Corthals, 2009, for review.) This approach limits the extent to which such measures can index quality variations in continuous speech rather than simply overall levels of dysphonia.

Previous studies comparing types of voice samples often, but not always (see, e.g., Awan et al., 2010; de Krom, 1994; Parsa & Jamieson, 2001; Revis et al., 1999), demonstrated differences in acoustic and perceptual measures of voice quality between the two stimulus types, with rating reliability usually in the low-to-moderate range, but equal for both kinds of stimuli (e.g., de Krom, 1994; Law et al., 2012; Maryn & Roy, 2012; Revis et al., 1999). The different kinds of information available from connected speech and steady-state vowels have led some authors to advocate inclusion of both kinds of stimuli in evaluations of voice quality (e.g., Maryn & Roy, 2012; Wolfe, Fitch, & Cornell, 1995), and the Acoustic Voice Quality Index (Maryn et al., 2010a; Maryn, De Bodt, & Roy, 2010b) and CAPE-V protocols, in fact, include analyses of both continuous speech and sustained vowels.

Although attention to both sustained vowels and connected speech during clinical evaluation seems a reasonable course of action given the different kinds of information available from different stimuli, questions of equivalence remain of concern to both researchers and clinicians, as evidenced by the continuing appearance of studies on the topic. One possible reason for this ongoing interest may be the difficulty of comparing the information conveyed by such divergent stimuli. In a sense, asking whether vowels provide all the information that speech does is a comparison of apples to oranges, and results will vary depending on whether the interest is primarily diagnostic (e.g., determining type or severity of a voice disorder) or if the interest is with developing a more complete understanding of normal and abnormal voice quality via well-controlled psychoacoustic studies. An alternative approach taken in the present study is to ask whether sustained vowels convey the same acoustic information contained in vowels excerpted from continuous speech and thus evoke the same perception of voice quality. If differences between voice quality measured from continuous speech and sustained vowels arise primarily from articulatory and prosodic influences that occur in continuous speech, then acoustic and perceptual measures of voice quality derived from vowel segments excised from continuous speech should not differ from those derived from sustained, steady-state vowels.

Method

Voice Samples

The voices of 10 speakers (six men and four women) with voice disorders were selected from a library of samples recorded under identical conditions. Samples were directly digitized at 20 kHz by using a Brüel and Kjær ½-in. microphone (model 4193; Nærum, Denmark), placed 10 cm from the speaker's lips at a 45° angle, with 16-bit resolution. No attempt was made to select stimuli that possessed any particular quality, except that the voice samples were chosen to range from normal to severely deviant.

Two types of vowel productions were elicited during each recording session. Speakers first sustained two repetitions each of the point vowels /i, u, a/ and then read three sentences with these vowels in stressed positions (see Table 1). A 1-s segment was excerpted from the middle of each sustained vowel. Steady-state portions of the target stressed vowels in the sentences (shown in bold type in Table 1) were excerpted by hand using custom software, with reference to both spectrograms and waveforms. Care was taken to avoid formant transitions so that the tokens were as steady in quality as possible (as verified by careful listening and examination of spectrograms by the second and third authors, who are both phoneticians). The two longest tokens of each of the three vowels /i, u, a/ were selected for analysis. Vowels excerpted from continuous speech had a mean duration of 145 ms (range = 59–270 ms; SD = 53 ms). Because these tokens were necessarily much shorter in duration than those available from sustained vowels, we created a third set of stimuli by further editing the sustained productions to match the duration of each excerpted vowel from the continuous speech samples.

Acoustic Evaluation

To determine whether acoustic parameters associated with voice quality varied with production context, measures

Table 1. Stimulus sentences.

No.	Sentence
1	Sh**e** s**ee**s a h**o**t sp**o**t.
2	The st**ew** is t**oo** h**o**t.
3	The s**ea** is full of t**u**na, c**o**d, and l**o**bster.

Note. Bold vowels were excerpted for analysis.

of the source spectral slope and of the noise-to-harmonics ratio (NHR) were made directly from the source spectrum for both sets of vowel recordings by using the analysis-by-synthesis method described by Kreiman, Antoñanzas-Barroso, and Gerratt (2010). In brief, the harmonic voice source was first estimated by inverse filtering the complete voice signal, using the method described by Javkin, Antoñanzas-Barroso, and Maddieson (1987).[1] The source spectrum was then smoothed by fitting it with a four-piece model whose segments ranged from the first to the second harmonic (H1–H2), from H2 to the harmonic nearest 2 kHz (H2–2 kHz), from the harmonic nearest 2 kHz to that nearest 5 kHz (2–5 kHz), and from H2 to the harmonic nearest 5 kHz (H2–5 kHz). These segments were chosen because they capture most of the variability in source spectral shapes (Kreiman, Gerratt, & Antoñanzas-Barroso, 2007a), their individual perceptual importance has been established (Garellek, Keating, Esposito, & Kreiman, 2013; Kreiman & Garellek, 2011), and in combination, they appear to form an adequate psychoacoustic model of source contributions to voice quality (Garellek, Samlan, Gerratt, & Kreiman, 2016; Kreiman, Garellek, Chen, Alwan, & Gerratt, 2015; Kreiman, Gerratt, Garellek, Samlan, & Zhang, 2014). These measures were thus preferred to others found in the literature (jitter, shimmer, etc.) whose relationship to overall voice quality is unknown. All variations in harmonic amplitudes within the specified ranges were eliminated so that the spectrum decreased smoothly in a piecewise fashion.

The spectral characteristics of the inharmonic part of the source (the noise excitation) were estimated using cepstral domain analysis similar to that described by de Krom (1993). These source estimates were then imported into a custom speech synthesizer (described in detail in Kreiman et al., 2010), and the harmonic amplitudes, NHR, and formant frequencies and bandwidths were adjusted until the synthetic voice formed a very close spectral and perceptual match to the target natural voice sample (see Figure 1) such that none of the authors could confidently distinguish the original voice sample from the synthetic copy. At this point, the spectral slope and NHR measures were recorded.

Perceptual Evaluation

To complement the acoustic analyses, we assessed the perceptual equivalence of the different stimuli in the following experiment, in which 12 listeners (seven men and five women, aged 19–32 years; mean age = 22.9, *SD* = 4.3) participated. All listeners reported normal hearing. All procedures were approved by the University of California, Los Angeles, Institutional Review Board.

To avoid measurement errors related to partitioning quality into components such as breathiness or roughness (Kreiman, Gerratt, & Ito, 2007b), listeners were asked to place each stimulus along a continuum from *normal* to *severely pathological* in a visual sort-and-rate task implemented in PowerPoint (Esposito, 2010; Granqvist, 2003). This procedure is quite similar to direct magnitude estimation but without the problem of drift in scale values due to memory limitations (because listeners can hear and compare any of the stimuli while they scale them; note that virtually all previous studies comparing vowels and continuous speech also used judgments of overall severity of pathology; see Maryn et al., 2010a, for review). Each listener participated in six trials, each including 30 stimuli (three vowels × 10 speakers). Two trials included 1-s sustained vowels, two included shortened sustained vowels, and two included vowels excerpted from continuous speech. Each of the two trials in a category included different vowel tokens, so each voice sample was judged only once by each listener. All stimuli in a trial were simultaneously presented on a computer screen as small icons of different shapes and colors (see Figure 2). Subjects listened to each stimulus by clicking its icon. If the stimulus sounded normal, listeners were instructed to drag its icon into a box labeled *normal voices* at the right side of the screen. They were asked to arrange the remaining voices from right to left in better-to-worse order along a 20-cm line extending left from the box on the screen such that the distance between icons represented the perceptual difference in severity between the voices. The listeners also were instructed that if two stimuli were judged to be equally severe, they should be stacked at the same position along the line. Listeners could play each stimulus and adjust its position on the line as often as needed, in any order, until they were satisfied with their sort.

Trials were randomized for each listener. Listeners were tested individually in a sound-treated room. They were encouraged to take breaks as needed. Stimuli were presented at a comfortable listening level (about 75 dB SPL) over Etymotic ER-1 insert earphones (Etymotic Research Inc., Elk Grove Village, IL), which mimic free-field presentation. Testing time averaged about 60 min.

After testing, the PowerPoint slides were printed, and listener responses were measured to the nearest 0.1 mm with a digital caliper, as the distance of each symbol from the left edge of the 20-cm line. All voices placed in the *normal* box were assigned a score of 20. Twelve percent of ratings were independently remeasured by a second lab assistant to assess the reliability of this procedure; the mean absolute difference between the first and second ratings was 0.14 mm (*SD* = 0.12 mm). Because listeners differed in how much of the line they used in a trial, responses for each stimulus series for each listener were normalized to a range of 0%–100%.

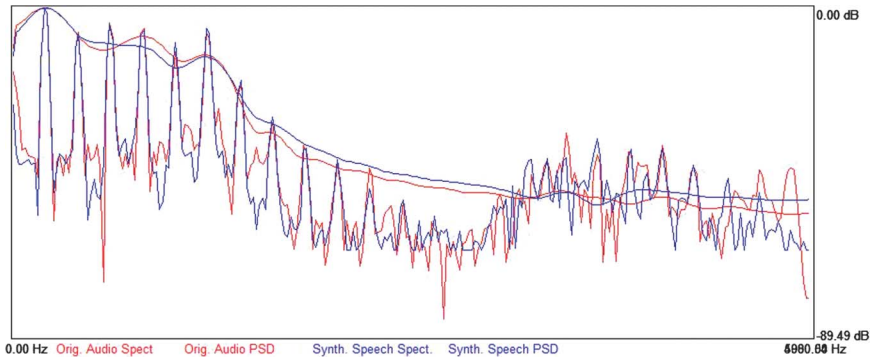
Results

Acoustic Measurements

Because acoustic measures were moderately intercorrelated (H1–H2 and H2–2 kHz, $r = -.34$; H2–2 kHz

[1]Source spectral measures also can be estimated from the complete speech signal (Hanson, 1997; Hanson & Chuang, 1999), although this procedure is somewhat prone to errors due to formant and bandwidth misestimation.

Figure 1. Spectra of an original natural voice source (red) and its synthetic copy (blue).



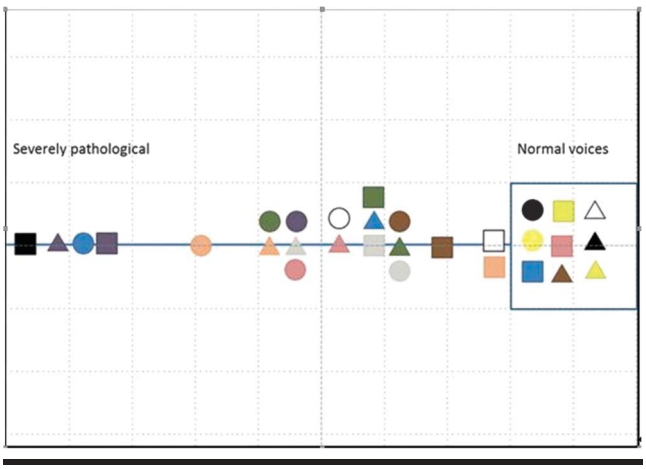
and H2–5 kHz, $r = .66$; H2–2 kHz and 2–5 kHz, $r = .59$; all $p < .01$ after Bonferroni adjustment), acoustic differences between shortened sustained vowels and vowels excerpted from continuous speech were assessed with five separate two-way (vowel by continuous/steady-state) analyses of variance (ANOVAs; with Bonferroni correction for multiple comparisons), one for each acoustic measure. Stimulus duration was not included as a factor in these analyses: Because the shortened vowels were excerpted from the sustained vowels, acoustic measures were essentially identical for these stimuli. No significant differences were observed between the sustained and excerpted vowels for any measure (see Table 2).

Listener Ratings

Overall Differences Across Listeners

A one-way ANOVA, with mean vocal severity rating as the dependent variable and listener as the independent variable, showed a significant main effect of listener, $F(11, 2148) = 17.24$, $p < .0001$. Post hoc t tests (with Bonferroni corrections) revealed that listeners could be divided into two groups that differed significantly in mean rating: one group, including three of the 12 listeners (mean severity = 0.48), with the remaining nine forming the second group (mean severity = 0.65). However, the main effect of listener accounted for only a small amount of variance ($\eta^2 = .081$). Therefore, in subsequent analyses, we grouped all listeners together, providing the results for the listener subgroups in footnotes.

Figure 2. Sample results from the sort-and-rate task.



Effect of Stimulus Length on Vocal Severity of Sustained Vowels

Next, we determined whether the length of the sustained vowel tokens affected perceived vocal severity. Recall that sustained vowels were presented either as 1-s samples or as shortened samples so that they could be compared more fairly to the excerpted tokens (see the "Comparison Between Shortened Sustained and Excerpted Tokens" section). A one-way ANOVA, with mean vocal severity rating as the dependent variable and length of the sustained vowel (long vs. shortened) as the independent variable, showed a small but significant main effect of length, $F(1, 1416) = 7.48$, $p < .05$. A post hoc t test revealed that, on average, the shortened sustained vowels were perceived as significantly less severely disordered than their long counterparts. This difference, though significant, was small (nine out of a maximal normalized rating range of 100) and accounted for very little variance in the rating data ($\eta^2 = .02$). Shortened sustained tokens had a mean vocal severity rating of 0.65, compared with 0.56 for long sustained tokens, indicating that both had mean ratings in the middle range of the scale.[2]

Comparison Between Shortened Sustained and Excerpted Tokens

To assess whether listeners rated excerpted and shortened sustained vowels differently, we ran a two-way ANOVA

[2]When the listeners were separated into two subgroups, the separate ANOVAs revealed that the smaller subgroup (with three listeners) was responsible for the effect of length on the mean rating for sustained vowels, $F(1, 354) = 337.92$, $p < .01$; $\eta^2 = .03$ (see Table A1). Listeners in the larger subgroup showed no effect of length on rated severity, $F(1, 1062) = 3.46$, $p > .1$.

Table 2. Mean values and analysis of variance results for comparisons of acoustic measures taken from sustained vowels versus continuous speech.

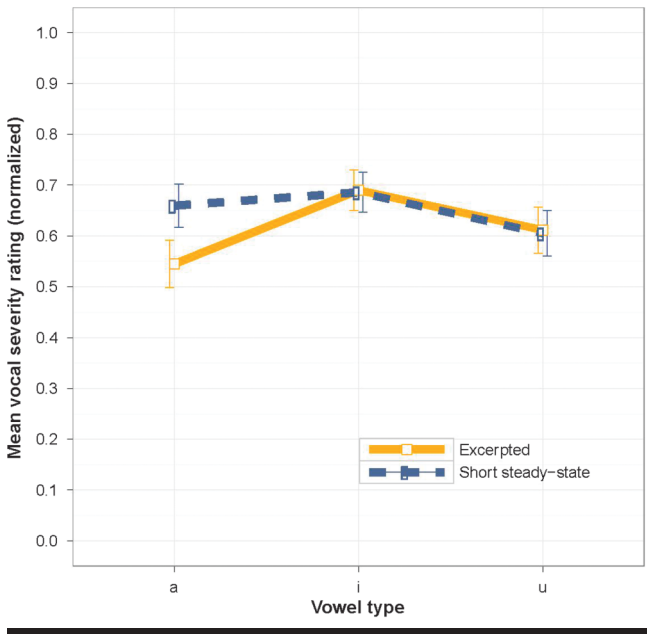
Acoustic measure	Mean (sustained; dB)	Mean (continuous; dB)	*F* value	*p* value
NHR	–28.29	–26.42	1.87	>.01
H1–H2	12.57	10.21	4.48	>.01
H2–2 kHz	28.13	27.36	0.24	>.01
2–5 kHz	8.54	8.39	0.01	>.01
H2–5 kHz	36.62	35.75	0.25	>.01

Note. Degrees of freedom are (1, 114) for all tests. No tests were significant after Bonferroni corrections for multiple comparisons. NHR = noise-to-harmonics ratio; H1 = first harmonic; H2 = second harmonic.

with listener rating of vocal severity as the dependent variable and token type (sustained vs. excerpted) and vowel (/i, u, a/) as the independent variables. There was a significant interaction between vowel and token type, $F(2, 1368) = 9.88$, $p < .001$, a significant main effect of vowel, $F(2, 1368) = 5.73$, $p < .01$, but no significant main effect of token type, $F(1, 1368) = 1.71$, $p > .05$ (see Figure 3).

Pairwise post hoc tests (with Bonferroni corrections) revealed that the main effect of vowel was driven by /i/, which was perceived as significantly less severe overall than /u, a/ ($p < .0001$), and that the interaction between vowel and token type was driven by /a/, for which sustained tokens had milder ratings than excerpted ones ($p < .01$). The effect sizes for both the main effect and interaction were small: The main effect of vowel type accounted for 1.3% of the overall variance, and the interaction between vowel and token type accounted for only 0.7% of the overall variance.[3]

Figure 3. Mean rating of vocal severity as a function of vowel (/a, i, u/) and token type (excerpted vs. sustained). Higher ratings indicate less severe-sounding voices. Error bars indicate 95% confidence intervals.



Discussion

Overall, this study demonstrated that sustained vowels and vowels excerpted from continuous speech contain essentially the same acoustic and perceptual information about extent of deviation from normal voice quality. No significant differences in acoustic measures of source characteristics were observed for the two kinds of stimuli, and no main effect of stimulus type was observed. Significant effects of stimulus length and vowel occurred, as did differences among listeners and an interaction between vowel and stimulus type, but these accounted for such small amounts of variance in the data that their practical importance is negligible. Because no meaningful acoustic or perceptual differences between stimulus types remained after controlling for variability related to continuous speech, these results imply that previously reported perceived and/or measured differences between continuous speech and steady-state vowels derive largely from source variability across segmental and prosodic contexts rather than from differences in patterns of vocal fold vibration in the quasisteady portion of the vowels.

If differences between stimulus types represent the presence versus absence of dynamic changes in quality rather than differences in the nature of the voice qualities themselves, this leads us to a paradox: Because most studies of connected speech apply measures that produce a single value for whole utterances, they do not actually assess the variations in quality that separate continuous from steady-state phonation, such variations being the ostensible reason for studying continuous speech in the first place. Instead,

[3]Separate ANOVAs for the two listener subgroups showed that the larger subgroup was responsible for the significant main effect of vowel type, $F(2, 1026) = 16.53$, $p < .001$; $\eta^2 = .029$, and the significant interaction between vowel type and token type, $F(2, 1026) = 6.94$, $p < .01$; $\eta^2 = .006$. The smaller subgroup showed no significant main effects or interaction: main effect of vowel, $F(2,342) = 0.88$, $p > .05$; main effect of token type, $F(1,342) = 7.87$, $p > .05$; interaction term, $F(2, 342) = 3.38$, $p > .05$ (see Table A2 for mean values associated with each statistical test).

perceptual studies usually assess the overall extent of dysphonia, as noted previously, while studies examining correlations between acoustic and perceptual measures usually use means and/or standard deviations for acoustic measures calculated across the entire sample of speech (e.g., Awan et al., 2009; de Krom, 1994; Halberstam, 2004; Lederle et al., 2012; Lowell et al., 2011; Maryn et al., 2010a; Maryn & Roy 2012; Moon et al., 2012; Parsa & Jamieson, 2001; Revis et al., 1999; Watts & Awan, 2011; Zraick et al., 2005; see Lowell, 2012, or Maryn et al., 2009, for review.) We know of only partial exceptions to this rule; for example, Watts and Awan (2011) found reliable differences in fundamental frequency and in the "cepstral spectral index of dysphonia" made over the first versus the second half of an eight-word sentence, presumably representing a confound of prosodic and pathologic factors (see also Stráník, Čmejla, & Vokřál, 2014). However, the changes in phonation associated with articulatory context can be very short, lasting as little as 25 ms (three to five cycles; Blankenship, 2002) or as long as about half of the vowel duration (Garellek, 2012). Measurement approaches with longer windows would not capture such rapid changes (although in some cases, such measures may correlate well with measures of perceived quality from long utterances; see, e.g., Hillenbrand and Houde, 1996, for the study of breathiness and the smoothed cepstral peak prominence [CPPs].

Complications deepen when we consider how one might better quantify the variations in voice quality in continuous speech. One solution might be to treat voice quality in connected speech as a concatenation of steady states, but this solution opens a Pandora's box of other complications. For example, in selecting the concatenating segments, how do we separate aspects of the speech signal that belong to voice quality and those that appertain to adjacent consonants? Consider, for example, the relationship between voicing onset following the aspirated release of /p/ or /t/ in English. When errors occur in voicing onset (say, aspiration is a bit prolonged), is this a voice disorder or a problem with articulatory timing related to the consonant? Is it necessary to include a model of articulation as part of our model of voice quality? How do we quantify (or even qualify) the different patterns of changes in voice that occur across different contexts, across different utterances, or across different moods, levels of fatigue, and speaking rates? And how do we compare measures from different utterances and situations once we have decided how to measure?

These points suggest that those favoring continuous speech for its ecological validity may pay a high price in terms of the internal validity of the quality measurements they make, to the point that it may no longer be clear what is being measured. Concerns about internal and external validity often occur in experimental design because of their inherently conflicting nature (Campbell & Stanley, 1963). For example, a gain in experimental control increases internal validity but may simultaneously reduce ecological and/or external validity. Sustained vowels by nature have far less variability in vocal amplitude, frequency, and quality, and therefore reduced psychoacoustic complexity, compared with segments of continuous speech, and also provide researchers and clinicians with a greater level of control over nonsource-related acoustic variability. However, the study of continuous speech segments allows measurement of voice occurring in more realistic natural speech contexts, in which the voice may vary in ways that are characteristic of a particular disorder. For example, the voice breaks and fluctuating voice quality that occur in adductory spasmodic dysphonia are much more readily observed during continuous speech. Because of this conflict, it seemingly follows that the selection of the kind of speech sample on which to base measurements or judgments of voice quality rests primarily on the research or clinical purpose. Current approaches using continuous speech stimuli, unfortunately, do not adequately quantify the acoustic and/or perceptual variability they are intended to assess, while simultaneously introducing serious challenges to internal measurement validity. In other words, the idea of measuring quality from continuous speech would be a good one, if only we knew what to measure, how to measure, and how to compare measurements across contexts.

Psychoacoustic studies have shown that listener agreement increases when the stimuli are carefully controlled to reduce their acoustic complexity (e.g., Kreiman et al., 2007b; Kreiman & Gerratt, 2011). Rating validity depends, in part, on reliability (Kreiman & Gerratt, 1998): Poor interrater agreement implies that ratings from different listeners are not comparable, so when listener disagreement in voice quality assessment is high, the internal validity of its measurement is not supported. Thus, researchers and clinicians who are interested in greater control over the acoustic stimulus will often study sustained vowels, because these relatively stationary stimuli foster greater listener agreement associated with their reduced acoustic complexity and thus more straightforward, psychoacoustic interpretation of listener responses (e.g., Kreiman et al., 2007b). Increased control of stimuli also reduces the sources of variability related to the difficulties listeners experience in isolating individual acoustic–perceptual or temporal attributes in complex voice patterns, leading to increased intra- and interlistener agreement and less difficulty in the interpretation of results (Kreiman et al., 2007b). Thus, the steady-state portions of any vowels as stimuli are more suited for carefully controlled psychoacoustic studies for the future development of a broader theoretical understanding of voice quality. As these psychoacoustic models of steady-state phonation develop to account for more and more aspects of quality, they also are likely to provide a foundation for the description and understanding of time-varying aspects of phonation. In this way, models of the voice quality expressed in continuous speech remain an essential and achievable goal for future study.

Acknowledgments

This research was supported by Grant DC01797 from the National Institute on Deafness and Other Communication Disorders, awarded to UCLA. We thank Marisa Tice for extensive help with

acoustic analysis and synthesis and Norma Antoñanzas-Barroso for ongoing programming support. Software used in this study is available without charge by request to the first or second author.

References

Awan, S. N., Roy, N., & Dromey, C. (2009). Estimating dysphonia severity in continuous speech: Application of a multi-parameter spectral/cepstral model. *Clinical Linguistics & Phonetics, 23,* 825–841.

Awan, S. N., Roy, N., Jette, M. E., Meltzner, G. S., & Hillman, R. E. (2010). Quantifying dysphonia severity using a spectral/cepstral-based acoustic index: Comparisons with auditory-perceptual judgements from the CAPE-V. *Clinical Linguistics & Phonetics, 24,* 742–758.

Blankenship, B. (2002). The timing of nonmodal phonation in vowels. *Journal of Phonetics, 30,* 163–191.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Boston, MA: Houghton Mifflin.

de Krom, G. (1993). A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals. *Journal of Speech and Hearing Research, 36,* 254–266.

de Krom, G. (1994). Consistency and reliability of voice quality ratings for different types of speech fragments. *Journal of Speech and Hearing Research, 37,* 985–1000.

Esposito, C. M. (2010). The effects of linguistic experience on the perception of phonation. *Journal of Phonetics, 38,* 306–316.

Garellek, M. (2012). The timing and sequencing of coarticulated non-modal phonation in English and White Hmong. *Journal of Phonetics, 40,* 152–161.

Garellek, M., Keating, P., Esposito, C. M., & Kreiman, J. (2013). Voice quality and tone identification in White Hmong. *The Journal of the Acoustical Society of America, 133,* 1078–1089.

Garellek, M., Samlan, R., Gerratt, B. R., & Kreiman, J. (2016). Modeling the voice source in terms of spectral slopes. *The Journal of the Acoustical Society of America, 139,* 1404–1410.

Granqvist, S. (2003). The visual sort and rate method for perceptual evaluation in listening tests. *Logopedics Phoniatrics and Vocology, 28,* 109–116.

Halberstam, B. (2004). Acoustic and perceptual parameters relating to connected speech are more reliable measures of hoarseness than parameters relating to sustained vowels. *ORL: Journal for Oto-Rhino-Laryngology, Head and Neck Surgery, 66,* 70–73.

Hanson, H. M. (1997). Glottal characteristics of female speakers: Acoustic correlates. *The Journal of the Acoustical Society of America, 101,* 466–481.

Hanson, H. M., & Chuang, E. S. (1999). Glottal characteristics of male speakers: Acoustic correlates and comparison with female data. *The Journal of the Acoustical Society of America, 106,* 1064–1077.

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

Javkin, H., Antoñanzas-Barroso, N., & Maddieson, I. (1987). Digital inverse filtering for linguistic research. *Journal of Speech and Hearing Research, 30,* 122–129.

Kempster, G. B., Gerratt, B. R., Verdolini Abbott, K., Barkmeier-Kraemer, J. M., & Hillman, R. E. (2009). Consensus Auditory-Perceptual Evaluation of Voice: Development of a standardized clinical protocol. *American Journal of Speech-Language Pathology, 18,* 124–132.

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

Kreiman, J., & Garellek, M. (2011). Perceptual importance of the voice source spectrum from H2 to 2 kHz. *The Journal of the Acoustical Society of America, 130,* 2570.

Kreiman, J., Garellek, M., Chen, G., Alwan, A., & Gerratt, B. R. (2015). Perceptual evaluation of voice source models. *The Journal of the Acoustical Society of America, 138,* 1–10.

Kreiman, J., & Gerratt, B. R. (1998). Validity of rating scale measures of voice quality. *The Journal of the Acoustical Society of America, 104,* 1598–1608.

Kreiman, J., & Gerratt, B. R. (2011). Comparing two methods for reducing variability in voice quality measurements. *Journal of Speech, Language, and Hearing Research, 54,* 803–812.

Kreiman, J., Gerratt, B. R., & Antoñanzas-Barroso, N. (2007a). Measures of the glottal source spectrum. *Journal of Speech, Language, and Hearing Research, 50,* 595–610.

Kreiman, J., Gerratt, B. R., Garellek, M., Samlan, R., & Zhang, Z. (2014). Toward a unified theory of voice production and perception. *Loquens, 1,* e009. doi:10.3989/loquens.2014.009

Kreiman, J., Gerratt, B. R., & Ito, M. (2007b). When and why listeners disagree in voice quality assessment tasks. *The Journal of the Acoustical Society of America, 122,* 2354–2364.

Law, T., Kim, J. H., Lee, K. Y., Tang, E. C., Lam, J. H., van Hasselt, A. C., & Tong, M. C. (2012). Comparison of rater's reliability on perceptual evaluation of different types of voice sample. *Journal of Voice, 26,* 666.e13–666.e21.

Lederle, A., Barkmeier-Kraemer, J., & Finnegan, E. (2012). Perception of vocal tremor during sustained phonation compared with sentence context. *Journal of Voice, 26,* 668.e1–668.e9.

Lowell, S. Y. (2012). The acoustic assessment of voice in continuous speech. *Perspectives on Voice and Voice Disorders, 22,* 57–63.

Lowell, S. Y., Colton, R. H., Kelley, R. T., & Hahn, Y. C. (2011). Spectral- and cepstral-based measures during continuous speech: Capacity to distinguish dysphonia and consistency within a speaker. *Journal of Voice, 25,* e223–e232.

Maryn, Y., Corthals, P., Van Cauwenberge, P., Roy, N., & De Bodt, M. (2010a). Toward improved ecological validity in the acoustic measurement of overall voice quality: Combining continuous speech and sustained vowels. *Journal of Voice, 24,* 540–555.

Maryn, Y., De Bodt, M., & Roy, N. (2010b). The Acoustic Voice Quality Index: Toward improved treatment outcomes assessment in voice disorders. *Journal of Communication Disorders, 43,* 161–174.

Maryn, Y., & Roy, N. (2012). Sustained vowels and continuous speech in the auditory-perceptual evaluation of dysphonia severity. *Jornal da Sociedade Brasileira de Fonoaudiologia, 24,* 107–112.

Maryn, Y., Roy, N., De Bodt, M., Van Cauwenberge, P., & Corthals, P. (2009). Acoustic measurement of overall voice quality: A meta-analysis. *The Journal of the Acoustical Society of America, 126,* 2619–2634.

Moon, K. R., Chung, S. M., Park, H. S., & Kim, H. S. (2012). Materials of acoustic analysis: Sustained vowel versus sentence. *Journal of Voice, 26,* 563–565.

Parsa, V., & Jamieson, D. G. (2001). Acoustic discrimination of pathological voice: Sustained vowels versus continuous speech. *Journal of Speech, Language, and Hearing Research, 44,* 327–339.

Revis, J., Giovanni, A., Wuyts, F., & Triglia, J. M. (1999). Comparison of different voice samples for perceptual analysis. *Folia Phoniatrica et Logopedica, 51,* 108–116.

Roy, N., Gouse, M., Mauszycki, S. C., Merrill, R. M., & Smith, M. E. (2005). Task specificity in adductor spasmodic dysphonia versus muscle tension dysphonia. *The Laryngoscope, 115,* 311–316.

Stráník, A., Čmejla, R., & Vokřál, J. (2014). Acoustic parameters for classification of breathiness in continuous speech according to the GRBAS scale. *Journal of Voice, 28,* 653.e9–653.e17.

Watts, C. R., & Awan, S. N. (2011). Use of spectral/cepstral analyses for differentiating normal from hypofunctional voices in sustained vowel and continuous speech contexts. *Journal of Speech, Language, and Hearing Research, 54,* 1525–1537.

Wolfe, V., Fitch, J., & Cornell, R. (1995). Sentence/vowel correlation in the evaluation of dysphonia. *Journal of Voice, 9,* 297–303.

Zraick, R. I., Wendel, K., & Smith-Olinde, L. (2005). The effect of speaking task on perceptual judgment of the severity of dysphonic voice. *Journal of Voice, 19,* 574–581.

Appendix

Mean Ratings for Subgroups of Listeners

Table A1. Mean ratings (with standard deviations in parentheses) of vocal severity for long versus shortened vowels for the two listener groups. Standard deviations are given parenthetically.

Stimulus	Group 1 (*n* = 3)	Group 2 (*n* = 9)
Sustained vowels	0.53 (0.32)	0.69 (0.33)
Shortened sustained vowels	0.42 (0.31)	0.60 (0.35)

Table A2. Mean ratings (with standard deviations in parentheses) of vocal severity for different vowels, stimulus types, and listener groups.

	Shortened sustained vowels		Excerpted vowels	
Vowel	Group 1 (*n* = 3)	Group 2 (*n* = 9)	Group 1 (*n* = 3)	Group 2 (*n* = 9)
/a/	0.63 (0.34)	0.67 (0.33)	0.48 (0.37)	0.57 (0.36)
/i/	0.51 (0.28)	0.74 (0.29)	0.48 (0.30)	0.76 (0.29)
/u/	0.46 (0.31)	0.65 (0.35)	0.50 (0.34)	0.65 (0.36)