Modeling the voice source in terms of spectral slopes[a)]

Marc Garellek
Department of Linguistics, University of California, San Diego, 9500 Gilman Drive #0108, La Jolla, California 92023-0108, USA

Robin Samlan
Department of Speech, Language, and Hearing Sciences, University of Arizona, Tucson, Arizona 85721-0071, USA

Bruce R. Gerratt and Jody Kreiman[b)]
Department of Head and Neck Surgery, UCLA School of Medicine, Los Angeles, California 90095-1794, USA

(Received 16 February 2015; revised 24 December 2015; accepted 3 March 2016; published online 29 March 2016)

A psychoacoustic model of the voice source spectrum is proposed. The model is characterized by four spectral slope parameters: the difference in amplitude between the first two harmonics (H1–H2), the second and fourth harmonics (H2–H4), the fourth harmonic and the harmonic nearest 2 kHz in frequency (H4–2 kHz), and the harmonic nearest 2 kHz and that nearest 5 kHz (2 kHz–5 kHz). As a step toward model validation, experiments were conducted to establish the acoustic and perceptual independence of these parameters. In experiment 1, the model was fit to a large number of voice sources. Results showed that parameters are predictable from one another, but that these relationships are due to overall spectral roll-off. Two additional experiments addressed the perceptual independence of the source parameters. Listener sensitivity to H1–H2, H2–H4, and H4–2 kHz did not change as a function of the slope of an adjacent component, suggesting that sensitivity to these components is robust. Listener sensitivity to changes in spectral slope from 2 kHz to 5 kHz depended on complex interactions between spectral slope, spectral noise levels, and H4–2 kHz. It is concluded that the four parameters represent non-redundant acoustic and perceptual aspects of voice quality.

© *2016 Acoustical Society of America*. [http://dx.doi.org/10.1121/1.4944474]

[MAH] Pages: 1404–1410

I. INTRODUCTION

In a recent paper, Kreiman *et al.* (2014) proposed a psychoacoustic model designed to capture the relationship between the acoustic voice signal and overall perceived voice quality.[1] Such a model is an essential component of a theory linking voice production to perception, so that when changes in the vocal production system cause changes to the acoustic signal, the acoustic changes can be used to predict the change in voice quality. Such information could provide useful insight into the evaluation and treatment of voice disorders, and could lead to better understanding of speaker recognition.

For these predictive relationships to function as intended, it would be helpful for each parameter modeling the voice source to be independent of the others: the more independent the parameters, the more unambiguously a single set of model parameters can be associated with a given target voice quality. This paper reports a series of experiments designed to establish the acoustic and perceptual relationships among parameters in the proposed model of the voice source.

A. The voice quality model

The psychoacoustic model of voice quality described by Kreiman *et al.* (2014) includes parameters to model the harmonic source spectrum, the inharmonic (noise) source, temporal source frequency ($F0$) and intensity characteristics, and the vocal tract transfer function. The perceptual importance of $F0$ (perceived as pitch; e.g., Moore, 1973) and amplitude (loudness; Stevens, 1936; see, e.g., Fastl and Zwicker, 2007, for review) as vocal attributes is well known, and listener sensitivity to changes in formant frequencies and bandwidths has also been extensively studied (e.g., Flanagan, 1955, 1957a). In this paper, we therefore focus on the shape of the harmonic voice source spectrum and its perceptual interactions with the inharmonic (noise) part of the complete voice source.

Our proposed source spectral model quantifies the spectrum with four parameters: the differences in dB between the amplitudes of the first two harmonics (H1–H2), the second and fourth harmonics (H2–H4), the fourth harmonic and the harmonic nearest 2 kHz in frequency (H4–2 kHz), and the harmonic nearest 2 kHz and that nearest 5 kHz (2 kHz–5 kHz). Within each band, harmonic amplitudes are set to decrease smoothly (Fig. 1). Thus, we assume that only the overall spectral shape, and not its fine details, is perceptually important. Moreover, the model implies that amplitudes of the lowest four harmonics must be relatively accurate, while individual harmonic amplitudes above H4 are much less relevant to perceived voice quality.

These particular parameters were chosen based on a principal component analysis of the spectra of 60 pathologic and 10 normal voices (28 males, 42 females; Kreiman *et al.*,

[a)]Portions of this paper were presented at the 165th and 166th Meetings of the Acoustical Society of America, 2013.
[b)]Electronic mail: jkreiman@ucla.edu

2007), which indicated that H1–H2, the spectral slope in the mid-frequencies (roughly 1–3 kHz), and high-frequency excitation (harmonic and inharmonic) together accounted for more than 88.5% of the variance in spectral shapes across talkers. Consideration of correlations between these components and other acoustic measures of spectral shape led to selection of H1–H2 and H2–H4 to model the lower part of the spectrum. Further examination of the source spectra indicated that "elbows"—abrupt changes in spectral slope—often occurred around 2 kHz (Kreiman *et al.*, 2011; Kreiman and Gerratt, 2011). As a result, two high-frequency component slopes, H4–2 kHz and 2 kHz–5 kHz, were included in the model (Garellek *et al.*, 2013b).[2]

Although the original set of source parameters was derived from principal components analysis, which yields orthogonal factors, the current set deviates sufficiently from those results that quasi-independence cannot be assumed. In addition, the extent of possible independence is constrained by the approximately −12 dB/octave decrease in source spectral energy reported by Flanagan (1957b). However, the extent to which naturally occurring spectral roll-off deviates from this ideal pattern, and the perceptual importance of any such deviations, is not currently known. For these reasons, three experiments were undertaken to assess acoustic and perceptual dependencies among model parameters. In experiment 1, we fit the model to a large number of empirically derived voice sources and examined the extent to which each component could be predicted acoustically from the others. Experiment 2 examined the perceptual independence of the source parameters by determining the extent to which changes in slope in one frequency range affected perceptibility of changes in another range. Finally, in experiment 3, we examined the perceptual dependence between spectral noise levels and harmonic energy above 2 kHz.

II. EXPERIMENT 1: CROSS-SPEAKER VARIABILITY IN SPECTRAL PROFILES

A. Method

Source spectra were measured for 144 voice samples produced by 136 individuals (74 female, 62 male), using analysis-by-synthesis (AbS) according to the method described in Kreiman *et al.* (2010). The original voices were one-second productions of /a/ recorded at 20 kHz using a Brüel and Kjær (Nærum, Denmark) 1/2 inch microphone. The voices in this study ranged from normal to severely disordered in quality, and a wide range of diagnoses were represented, including reflux laryngitis, mass lesions, and functional and neurogenic disorders.

In the first step of the AbS process, parameters describing the harmonic part of the voice source were estimated as follows. First, a representative cycle from each voice sample was inverse filtered (Javkin *et al.*, 1987). Twenty identical pulses were concatenated and the source spectrum was calculated by FFT from this series. Segments were selected for each of the four source parameters (H1–H2, H2–H4, H4–2 kHz, and 2 kHz–5 kHz), and all harmonic amplitudes within each range were adjusted so that the spectrum decreased smoothly within each segment (Fig. 1). The spectrum of the inharmonic part of the source (the noise excitation) was estimated using 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 $F0$ and amplitude contours, $F0$ was tracked pulse by pulse on the time domain waveform. Formant frequencies and bandwidths were estimated using autocorrelation linear predictive coding analysis with a window of 25.6 ms.

The synthesizer's sampling rate was fixed at 10 kHz. $F0$ and amplitude contours were applied by time and amplitude

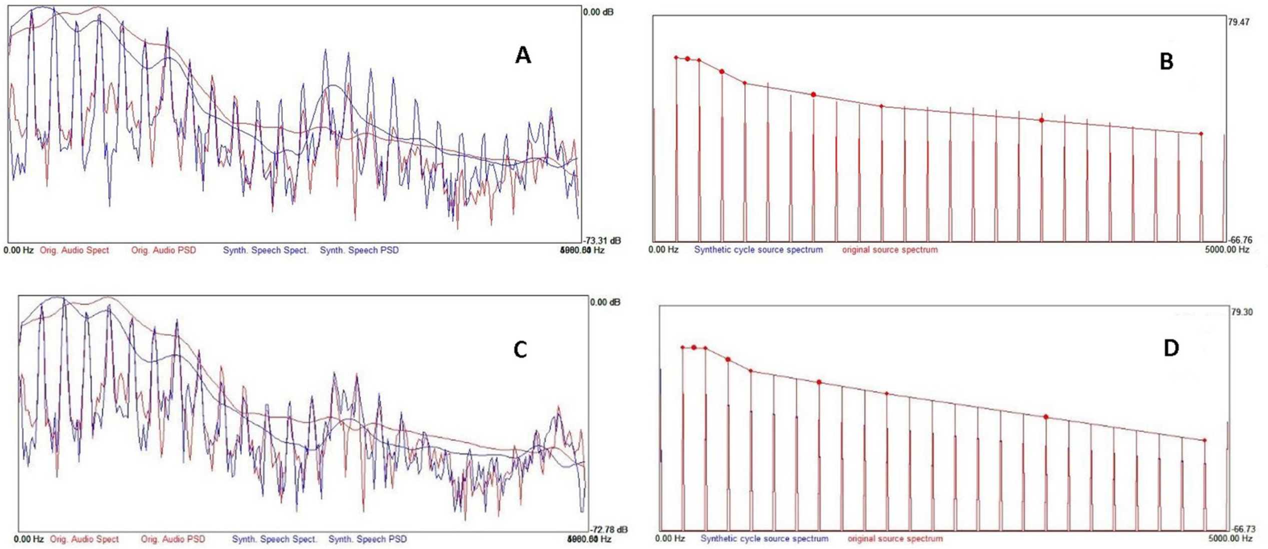

FIG. 1. (Color online) Spectra of representative original and synthetic voices, with the associated source spectra. (A) Voice spectra before AbS. Note mismatches between natural and synthetic spectra. (B) The source spectrum before source model fitting. Note excursions of individual harmonics above and below the line segments, indicating model parameters. (C) The voice spectrum after parameter adjustment. (D) The smoothed source spectrum after harmonic amplitude adjustment. Note that variability in the amplitude of individual harmonics has been eliminated (see Kreiman *et al.*, 2010, for details of this method).

warping individual source pulses and then concatenating them to form a complete source time series. Spectra remained constant across all pulses. The spectral noise time series was then added to the harmonic source, and the complete (harmonic + inharmonic) synthesized source was filtered through the vocal tract model. Finally, parameters were adjusted until the synthetic copy matched the target natural voice stimulus spectrally and perceptually, as judged by the authors. Although in theory ambiguity exists in this process (due to the fact that harmonic amplitudes can be modified by adjustments to either bandwidths or source characteristics), in practice, adjusting the slope of a source parameter affected the amplitude of a range of harmonics, while changes to bandwidths affected at most 1–2 harmonics, reducing ambiguity. Once the synthetic copy matched the target voice spectrally and perceptually, the four source spectral slope measures were recorded along with $F0$, the noise-to harmonics ratio (NHR), and the spectral slope from H1 to the highest harmonic (H1–5 kHz) as an estimate of overall spectral roll-off.

B. Results and discussion

1. Ranges of values for each component slope

Means, standard deviations, and ranges for each source spectral slope parameter are listed in Table I. For most voices, H1–H2 and H2–H4 values fell between 0 and 20 dB. Only four voices had H1–H2 values below 0 dB, and only one voice had a negative value for H2–H4. The higher-frequency component slopes had slightly larger ranges, with most values falling between 0 and 40 dB. No voices had a negative H4–2 kHz slope, and only three voices had negative slopes for 2 kHz–5 kHz.

2. Relationships among spectral slope components, F0, speaker sex, NHR, and H1–5 kHz

Linear regression was used to model H1–H2, H2–H4, H4–2 kHz, and 2 kHz–5 kHz, each as a function of the other parameters. Two analyses were performed for each spectral parameter: one including the other spectral components plus overall slope (H1–5 kHz), NHR, $F0$, speaker sex, and the interaction between $F0$ and speaker sex, and one including only significant predictors from among the latter set of variables. The difference in variance accounted for in the two analyses served as an estimate of the strength of the relationship between the different spectral model components after controlling for non-spectral parameters and overall spectral roll-off (as estimated by H1–5 kHz).

TABLE I. Means, standard deviations (SD), and ranges for the spectral slope components in dB, separated by speaker sex.

	Female voices		Male voices	
	Mean (SD)	Range	Mean (SD)	Range
H1–H2	8.93 (4.55)	−2.0–21.6	6.13 (4.11)	−0.7–21.5
H2–H4	11.57 (4.99)	2.0–29.2	8.93 (3.74)	−5.1–19.9
H4–2 kHz	18.08 (6.66)	2.0–37.7	24.58 (6.58)	12.2–43.2
2 kHz–5 kHz	16.20 (9.20)	−3.0–41.3	15.49 (8.23)	−3.8–38.5

Results are shown in Table II. H1–H2 was significantly predicted by H2–H4, H4–2 kHz, 2 kHz–5 kHz, NHR, and H1–5 kHz [$F(8,135) = 30.99$, $p < 0.0001$, adjusted $R^2 = 0.63$]: higher H1–H2 (a steeper spectral slope) corresponded to lower values of H2–H4, H4–2 kHz, and 2 kHz–5 kHz (all $p < 0.0001$), higher overall slope ($p < 0.0001$), and NHR ($p < 0.05$). The spectral slope components added 46% to the variance predicted by the smaller model (which excluded H2–H4, H4–2 kHz, and 2 kHz–5 kHz). H2–H4 was significantly predicted by H1–H2, H4–2 kHz, 2 kHz–5 kHz, $F0$, and overall slope [$F(8,135) = 38.31$, $p < 0.0001$, adjusted $R^2 = 0.68$]. Consistent with the trading relationship just described, higher values of H2–H4 were associated with lower values of H1–H2, H4–2 kHz, and 2 kHz–5 kHz (all $p < 0.001$), and higher values of overall slope ($p < 0.0001$) and $F0$ ($p < 0.05$). The spectral slope components added 44% to the variance predicted by the smaller model. H4–2 kHz was predicted by H1–H2, H2–H4, 2 kHz–5 kHz, $F0$, and overall slope [$F(8,135) = 92.06$, $p < 0.0001$, adjusted $R^2 = 0.84$], with higher values of H4–2 kHz being associated with lower values of H1–H2, H2–H4, and 2 kHz–5 kHz (all $p < 0.001$), lower values of $F0$ ($p < 0.05$), and higher values of overall slope ($p < 0.0001$). The spectral parameters added 45% to the explained variance. Finally, 2 kHz–5 kHz was predicted by H1–H2, H2–H4, 2 kHz–5 kHz, and overall slope [$F(8,135) = 118.50$, $p < 0.0001$, adjusted $R^2 = 0.87$], with higher values of H4–2 kHz being associated with lower values of H1–H2, H2–H4, and 2 kHz–5 kHz (all $p < 0.001$), and higher values of overall slope ($p < 0.0001$). The spectral parameters added 38% to the explained variance.

In summary, when overall spectral roll-off was controlled by including H1–5 kHz as a covariate in regression analyses, strong relationships among H1–H2, H2–H4, H4–2 kHz, and 2 kHz–5 kHz were observed; when overall spectral roll-off is not included in the regression analyses, H1–H2, H2–H4, H4–2 kHz, and 2 kHz–5 kHz are not otherwise individually predictable from each other (Garellek *et al.*, 2013b). In other words, dependencies among the spectral model parameters are a function of overall spectral roll-off: overall roll-off does not predict the value of any single spectral slope parameter very well, but it does explain the relationships among parameters, as indicated by the larger R^2 values in the first column of Table II. The value of each model component tended to have an inverse relationship with the slope of the adjacent segment(s), such that higher values of one slope (e.g., H1–H2) usually implied lower values of an adjacent slope (e.g., H2–H4), again consistent with the fact that the source spectrum rolls off with increasing

TABLE II. Variance predicted after controlling for non-model components.

Model component	R^2 for the full model	R^2 for the model including only NHR, $F0$, sex × $F0$, and H1–5 kHz	Difference
H1–H2	0.63	0.17	0.46
H2–H4	0.68	0.24	0.44
H4–2 kHz	0.84	0.39	0.45
2 kHz–5 kHz	0.87	0.49	0.38

frequency (Carr and Trill, 1964; Ní Chasaide and Gobl, 1997): the more one component of the model increases or decreases in slope, the more its neighbors must also change to maintain the overall decrease in spectral energy with increasing frequency.

III. EXPERIMENT 2: PERCEPTUAL INDEPENDENCE OF PARAMETERS OF THE HARMONIC SOURCE MODEL

The analyses in experiment 1 showed that components of the model of the harmonic voice source were correlated with adjacent components, but that the observed associations could be explained largely by overall spectral roll-off. Although as a result it may be possible to treat these parameters as acoustically independent within the constraints imposed by roll-off, this does not imply that components are perceptually independent. Although listeners are approximately equally sensitive to changes in the different spectral slope parameters (Kreiman *et al.*, 2014), the shape of the surrounding spectrum may foreground a given segment perceptually, increasing listeners' ability to discriminate changes in that parameter, so that sensitivity to a segment may depend on the shape of the surrounding spectrum. The following experiment examined this possibility.

A. Method

1. Stimuli

A synthetic copy of a 1 s sample of a normal-sounding female voice was created using the UCLA voice synthesizer (Kreiman *et al.*, 2010). For the original voice, H1–H2 equaled 7 dB, H2–H4 equaled 7 dB, H4–2 kHz equaled 15 dB, and 2 kHz–5 kHz equaled 7 dB. *F*0 equaled 200 Hz and the NHR equaled −29.7 dB.

Based on this synthetic sample, 12 blocks of stimuli were synthesized (Table III). Each block consisted of 30 stimuli varying on a single spectral slope parameter (e.g., H1–H2) with the slope of the adjacent component(s) (e.g., H2–H4) set at a fixed high or low value. The values for the adjacent components H1–H2, H2–H4, and H4–2 kHz were determined by adding or subtracting 1.5 times the just-noticeable-difference (JND) of that slope, as estimated in previous research (4.1 dB for H1–H2; 3.0 dB for H2–H4; 3.9 dB for H4–2 kHz; 11.5 dB for 2 kHz–5 kHz; Garellek *et al.*, 2013b). In the case of 2 kHz–5 kHz, the high value resulting from this formula exceeded the range observed for female voices (7 dB – 1.5 times a JND of 11.5 dB; Table I). As a result, values were chosen that were more consistent with the ranges shown in Table I.

TABLE III. Spectral slope components and their range of values for the stimuli in experiment 2.

Spectral slope component varying	Range of slopes for the varying component (dB)	Step size (dB)	Altered adjacent slope	Value of altered adjacent slope
H1–H2	4.0–18.5	0.5	H2–H4	11.5
				2.5
H2–H4	2.5–17	0.5	H1–H2	13.0
				1.0
			H4–2 kHz	21.0
				9.0
H4–2 kHz	10.0–24.5	0.5	H2–H4	11.5
				2.5
			2 kHz–5 kHz	36.0
				0.0
2 kHz–5 kHz	5.0–34	1.0	H4–2 kHz	21.0
				9.0

2. Participants and task

All experimental procedures were approved by the UCLA Institutional Review Board. Thirty-three listeners [UCLA students and staff; 23 female, mean age = 22 yr, standard deviation (SD) = 3.0] participated in the task. All reported normal hearing. They were compensated for their time.

Experiment 2 consisted of 12 blocks of stimuli (Table IV), as described above. An additional block was included to serve as pilot data for another study, and was discarded without analysis. Thirty-two listeners completed 6 of the 12 blocks (selected at random), and 1 listener completed only 3 blocks. In total, each block was judged by 15 listeners, and no listener heard the same block more than once.

Listeners were seated in a double-walled sound booth and heard the stimuli over Etymotic ER-1 insert earphones (Etymotic Research, Inc., Elk Grove Village, IL). On a given trial, listeners heard two stimuli separated by 250 ms of silence, and were asked to judge whether the two were the same or different. The first stimulus was always the first in the series (i.e., with the lowest slope value). The second stimulus differed only in the slope component being assessed; the initial value in a run differed from the first stimulus by seven steps. Listeners were able to play the two stimuli once in each order (AB and BA) before making their decisions. If the listener correctly distinguished the stimuli in two successive trials, then the difference between the stimuli was decreased by 0.5 dB (or 1.0 dB for stimuli where 2 kHz–5 kHz was varied). If the listener did not perceive the

TABLE IV. Summary results for experiment 2. The only significant effect was the increase in JND for 2 kHz–5 kHz when H4–2 kHz was steep (21 dB) vs flat (9 dB).

Target spectral slope component	Altered adjacent component	Adjacent slope (dB)	JND mean (SD) for target slope (dB)
H1–H2	H2–H4	11.5	5.71 (2.16)
		2.5	6.73 (2.72)
H2–H4	H1–H2	13.0	4.81 (2.03)
		1.0	4.91 (2.14)
	H4–2 kHz	21.0	5.97 (2.52)
		9.0	5.64 (2.61)
H4–2 kHz	H2–H4	11.5	4.24 (2.01)
		2.5	4.41 (2.63)
	2 kHz–5 kHz	36.0	6.31 (2.51)
		0.0	5.29 (2.09)
2 kHz–5 kHz	H4–2 kHz	21.0	23.81 (4.02)
		0.0	14.87 (6.20)

difference between the two stimuli, then the difference was increased by 0.5 dB (or 1.0 dB for stimuli where 2 kHz–5 kHz were varied). The run continued until 12 reversals were obtained, and the JND for each listener and block was calculated by averaging the difference between the standard and test stimuli for the last 8 reversals. This procedure identifies the spectral slope value for which the listener could correctly distinguish the target and test stimuli 70.7% of the time (Levitt, 1971).

B. Results and discussion

The JNDs for each spectral component are shown in Table IV. A one-way analysis of variance (ANOVA) indicated that the value of H2–H4 (high vs low) did not significantly affect the JND for H1–H2 [$F(1,25) = 0.22$, $p > 0.05$, $\eta^2 = 0.10$]. Neither the slope of H1–H2 nor H4–2 kHz significantly altered the JND for H2–H4 [$F(3,55) = 0.63$, $p > 0.05$, $\eta^2 = 0.07$], and neither the slope of H2–H4 nor 2 kHz–5 kHz significantly altered the JND for H4–2 kHz [$F(3,55) = 2.50$, $p > 0.05$, $\eta^2 = 0.24$].

For 2 kHz–5 kHz, 10 of the 15 listeners did not detect a difference in slope of 30 dB (the largest examined) between stimuli, and one listener reported hearing a difference between every pair of stimuli, so that JNDs could not be estimated. In these cases, we replaced "No JND" responses with either 30 dB (where listeners consistently heard no differences) or 1 dB (where the listener consistently heard a difference). A subsequent one-way ANOVA showed a significant effect of H4–2 kHz on sensitivity to changes in 2 kHz–5 kHz [$F(1,18) = 44.90$, $p < 0.05$]. A steep H4–2 kHz resulted in a significantly higher JND for 2 kHz–5 kHz than a flat profile of H4–2 kHz.

These results indicate that the parameters of the spectral source model are perceptually independent of adjacent model parameters, with the exception of 2 kHz–5 kHz. The interaction between H4–2 kHz and 2 kHz–5 kHz is difficult to interpret given the known interaction (described below) between spectral noise levels and listeners' sensitivity to changes in source spectral slope in the range from 2 kHz to 5 kHz. Experiment 3 examined the perceptual interactions among these parameters in greater detail.

IV. EXPERIMENT 3: LISTENER SENSITIVITY TO 2 kHz–5 kHz AS A FUNCTION OF NHR

Previous studies (Kreiman and Gerratt, 2005, 2012) have shown that perception of changes in the harmonic source spectral slope from the second harmonic to the harmonic nearest 5 kHz (H2–5 kHz) depended on NHR, on the harmonic source spectral slope, and on the shape of the noise spectrum: JNDs increased when spectra rolled off steeply, with this effect in turn depending on NHR level and spectral shape. (See also Shrivastav and Sapienza, 2003, 2006; Shrivastav and Camacho, 2010, for similar results derived from perceptual modeling of breathy voice quality.) However, the source spectral model used in those studies differed from the current one in the detail with which the higher frequencies were modeled (H2–5 kHz, vs H4–2 kHz and 2 kHz–5 kHz). Further, no associations between H4–2 kHz or 2 kHz–5 kHz and the NHR were found in experiment 1. This experiment extends our previous findings to the current harmonic source model and seeks to resolve the discrepancy in the results.

A. Method

The task was the same as in experiment 2. Fifteen listeners recruited from UCLA (12 female, mean age = 27.7 yr, SD = 5.9) participated. None of these listeners participated in experiment 2. All reported normal hearing. They were compensated for their time.

The stimuli in experiment 3 had the same values of 2 kHz–5 kHz as in the previous experiment (30 values varying by 1 dB increments). H4–2 kHz was held constant at 21 dB or 9 dB, and the other two component slopes were held constant at the values used in the previous experiment. NHR was increased to −10 dB (vs −29.7 dB in experiment 2).

B. Results and discussion

Detecting changes in high-frequency spectral slope in the context of high noise levels proved very difficult for listeners. When H4–2 kHz was set at 21 dB, 7 of the 15 participants did not converge on a JND for 2 kHz–5 kHz; and when H4–2 kHz was set at the flatter level of 9 dB, 9 of the 15 participants did not converge on a JND for 2 kHz–5 kHz.

These data were next combined with the JND data for 2 kHz–5 kHz from experiment 2. As before, a JND of 30 dB was assigned when listeners did not perceive a difference between stimuli that differed by this amount. Two-way between-subjects ANOVA examined the effects of NHR (high vs low) and H4–2 kHz slope (steep vs flat) and their interaction on JNDs for 2 kHz–5 kHz. Results revealed significant main effects of NHR [$F(1,59) = 5.11$, $p < 0.05$, partial $\eta^2 = 0.12$] and H4–2 kHz [$F(1,59) = 4.09$, $p < 0.05$, partial $\eta^2 = 0.12$], along with a significant interaction between NHR and H4–2 kHz [$F(1,59) = 9.64$, $p < 0.01$, partial $\eta^2 = 0.35$] (Fig. 2). Tukey pairwise comparisons revealed

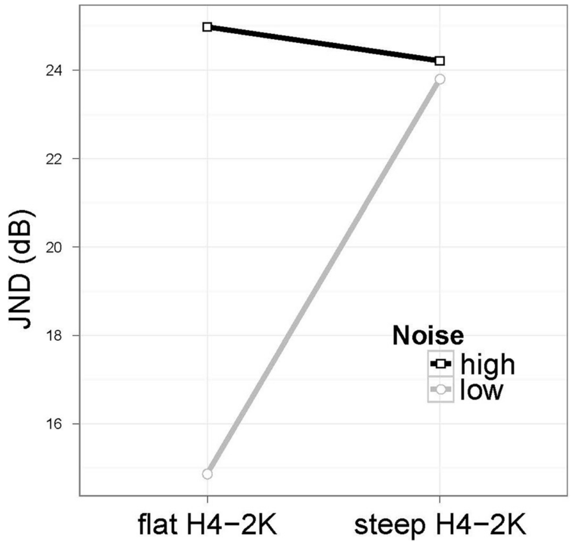

FIG. 2. Mean JND for 2 kHz–5 kHz as a function of H4–2 kHz slope and NHR.

that NHR only affected JNDs for 2 kHz–5 kHz when H4–2 kHz was flat ($p < 0.01$), and the slope of H4–2 kHz only affected the JND for 2 kHz–5 kHz when noise was low ($p < 0.01$; experiment 2).

These results are largely consistent with previous findings: changes in high-frequency spectral slope were harder to hear in the context of a large NHR than in the context of a low NHR, whether this slope parameter was defined as H2–5 kHz or as 2 kHz–5 kHz. However, this effect is mediated by the spectral slope from H4–2 kHz: when H4–2 kHz was relatively flat, JNDs for 2 kHz–5 kHz were smaller but dependent on spectral noise level, while a relatively steeply falling H4–2 kHz corresponded to higher JNDs for 2 kHz–5 kHz, regardless of noise level. Put another way, when H4–2 kHz is relatively flat, either changes in spectral slope or in noise levels will have an influence on voice quality, but when H4–2 kHz falls off more steeply, changes to frequencies above 2 kHz are relatively unimportant perceptually.

V. GENERAL DISCUSSION

The voice source spectrum can be partitioned in an infinite number of ways, of course, and many measures have previously been devised to quantify perceptually important spectral attributes (although, to our knowledge, no other psychoacoustic model of overall voice quality has been proposed). For example, authors have proposed measures of the deviation of the empirical source slope from an “ideal” slope in different frequency bands (typically, four bands, each 1 kHz wide, from 0 to 4 kHz; Sundberg and Gauffin, 1979; Ní Chasaide and Gobl, 1997), along with measures of the amplitude of H1 relative to that of the first and/or second formant (H1–A1 or H1–A2; Hanson, 1997), among others (see Kreiman *et al.*, 2007, for review). With this caveat, however, the results of the experiments reported here demonstrate that the three parameters modeling the spectral slope below 2 kHz are relatively independent, both acoustically and perceptually, and that listeners are approximately equally sensitive to all parameters. The observed patterns of trade-off in slope between spectrally adjacent parameters are consistent with spectral roll-off with increasing frequency. Further, sensitivity to H1–H2, H2–H4, and H4–2 kHz did not change as a function of the slope of an adjacent component, which suggests that sensitivity to these components is particularly robust.

Above 2 kHz, parameters are acoustically independent but listeners' sensitivity to changes in spectral slope from 2 kHz–5 kHz depends on a complex interaction between spectral slope, spectral noise levels, and the spectral slope between H4 and 2 kHz. We interpret this as follows. The lowest-frequency harmonics of the voice source (H1 through H4) carry the most information about overall glottal pulse shape/voice quality (e.g., Fant, 1995), with harmonics above H4 contributing relatively little information about overall pulse shape. Thus, perceptual and acoustic independence of the lowest-frequency harmonics is a *sine qua non* for a speaker's ability to produce a wide range of voice qualities across utterances and circumstances. Above H4, observed source spectra form a continuum in shape, with spectra that are flat from H4–2 kHz and falling from 2 kHz–5 kHz forming one endpoint and those that fall continuously from H4–5 kHz forming the other. Harmonic energy does not interact perceptually with spectral noise when spectra fall continuously (the second endpoint), because JNDs for harmonic energy levels are already very large in this frequency range. That is, when the spectral energy levels decrease steadily across frequencies, little energy is present in the high frequencies, so even small amounts of noise will form an effective masker. When the spectrum above H4 remains relatively flat, the voice has more high-frequency energy overall, which allows more possibility of interactions with spectral noise and smaller JNDs overall.

The relative independence of the four spectral parameters is consistent with previous work showing that they function contrastively to carry linguistic and other information. For example, listeners of various languages are sensitive to changes in H1–H2 (Esposito, 2010; Kreiman and Gerratt, 2012; Garellek *et al.*, 2013a), and H1–H2 is also known to vary according to differences in voice quality (Bickley, 1982; Klatt and Klatt, 1990; Gordon and Ladefoged, 2001). H2–H4 is related to the perception of pitch location within a speaker's range (Bishop and Keating, 2012), and its synthetic manipulation drives changes in the perception of contrastive breathy voice in Hmong (Garellek *et al.*, 2013a).

One important limitation of the present study is the use of a single voice as the frame within which the spectral slope components and NHR were varied. Because $F0$ appears to be a significant predictor of the slope of H1–H2, H2–H4, and H4–2 kHz, further work is required to determine the extent to which sensitivity to these component slopes varies (if it does) as a function of $F0$, sex, and additional model parameters. Although NHR was not a significant predictor of spectral slopes below 2 kHz, it is still possible that the noise level and/or the slope of the noise spectrum affects sensitivity to these slope components, and that spectral slope in turn influences sensitivity to NHR and other model parameters (Kreiman and Gerratt, 2012).

In conclusion, by modeling voice quality using a rather small set of perceptible parameters, we define it in a substantively different manner than do traditional paradigms using rating scales for attributes like roughness or breathiness (which are unreliable as measurement tools; Kreiman and Gerratt, 1998). This traditional atheoretical approach has, to date, failed to elucidate links between acoustics and either vocal physiology or perceived voice quality. In contrast, our proposed model specifies direct links between perception of overall quality and vocal acoustics, which may in turn make it possible to identify physiological changes that produce perceptually meaningful acoustic changes during voice production. Validation of this model and of this general approach to quality measurement is thus an important priority for voice research. In addition to providing a clearer definition of quality based on psychoacoustics rather than rating scale opinions, this model may allow a more direct understanding and prediction of how changes in vocal physiology produce an intended vocal quality to which listeners are perceptually sensitive, an issue that is particularly important in the evaluation and treatment of voice disorders.

ACKNOWLEDGMENTS

We are grateful to Norma Antoñanzas-Barroso for her assistance with the synthesizer and to Shaghayegh Rastifar for assistance with participants and data reduction. This research was supported by National Institutes of Health (NIH) Grant No. DC01797.

[1] In this paper, "quality" is used in the ANSI (1960) sense of those attributes that allow listeners to distinguish stimuli that sound different but are equal in pitch and loudness, and not in the sense of individual facets of quality like breathiness or hoarseness. See Kreiman and Sidtis (2011) for extended discussion.

[2] Subsequent experiments have also shown that perceptual modeling improves when the spectrum between H4 and 5 kHz is modeled in two pieces.

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

Bickley, C. (**1982**). "Acoustic analysis and perception of breathy vowels," MIT/RLE Work. Pap. Speech Commun. **1**, 71–82.

Bishop, J., and Keating, P. (**2012**). "Perception of pitch location within a speaker's range: Fundamental frequency, voice quality, and speaker sex," J. Acoust. Soc. Am. **132**, 1100–1112.

Carr, P. B., and Trill, D. (**1964**). "Long term larynx-excitation spectra," J. Acoust. Soc. Am. **36**, 2033–2040.

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.

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

Fant, G. (**1995**). "The LF-model revisited. Transformations and frequency domain analysis," Speech Trans. Lab. - Quart. Prog. Status Rep. **36**(2–3), 119–156.

Fastl, H., and Zwicker, E. (**2007**). *Psychoacoustics: Facts and Models* (Springer Science and Business Media, Berlin), pp. 1–462.

Flanagan, J. (**1955**). "A difference limen for vowel formant frequency," J. Acoust. Soc. Am. **27**, 613–617.

Flanagan, J. (**1957a**). "Difference limen for formant amplitude," J. Speech Hear. Disord. **22**, 205–212.

Flanagan, J. (**1957b**). "Note on the design of terminal-analog speech synthesizers," J. Acoust. Soc. Am. **29**, 306–310.

Garellek, M., Esposito, C. M., Keating, P., and Kreiman, J. (**2013a**). "Voice quality and tone identification in White Hmong," J. Acoust. Soc. Am. **133**, 1078–1089.

Garellek, M., Samlan, R. A., Kreiman, J., and Gerratt, B. R. (**2013b**). "Perceptual sensitivity to a model of the source spectrum," Proc. Meet. Acoust. **19**, 060157.

Gordon, M., and Ladefoged, P. (**2001**). "Phonation types: A cross-linguistic overview," J. Phonetics **29**, 383–406.

Hanson, H. M. (**1997**). "Glottal characteristics of female speakers: Acoustic correlates," J. Acoust. Soc. Am. **101**, 466–481.

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., 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., Garellek, M., and Esposito, C. (**2011**). "Perceptual importance of the voice source spectrum from H2 to 2 kHz," J. Acoust. Soc. Am. **130**, 2570.

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., and Gerratt, B. R. (**2011**). "Modeling overall voice quality with a small set of acoustic parameters," J. Acoust. Soc. Am. **129**, 2529.

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. (**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," Loquens **1**, e009.

Kreiman, J., and Sidtis, D. (**2011**). *Foundations of Voice Studies* (Wiley-Blackwell, Malden, MA), pp. 1–24.

Levitt, H. (**1971**). "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am. **49**, 467–477.

Moore, B. C. J. (**1973**). "Frequency difference limens for short-duration tones," J. Acoust. Soc. Am. **54**, 610–619.

Ní Chasaide, A., and Gobl. C. (**1997**). "Voice source variation," in *The Handbook of Phonetic Sciences*, edited by W. J. Hardcastle and J. Laver (Blackwell, Oxford), pp. 427–461.

Shrivastav, R., and Camacho, A. (**2010**). "A computational model to predict changes in breathiness resulting from variations in aspiration noise level," J. Voice **24**, 395–405.

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

Stevens, S. S. (**1936**). "A scale for the measurement of a psychological magnitude: Loudness," Psychol. Rev. **43**, 405–416.

Sundberg, J., and Gauffin, J. (**1979**). "Waveform and spectrum of the glottal voice source," in *Frontiers of Speech Communication Research*, edited by B. Lindblom and S. Ohman (Academic, London), pp. 301–322.