Research Article

Interactions Between Breathy and Rough Voice Qualities and Their Contributions to Overall Dysphonia Severity

Yeonggwang Park,[a] Supraja Anand,[a] Lisa M. Kopf,[b] Rahul Shrivastav,[c,d] and David A. Eddins[a,d]

[a] Department of Communication Sciences and Disorders, University of South Florida, Tampa [b] Department of Speech, Language and Hearing Sciences, The George Washington University, Washington, DC [c] Office of the Provost and Executive Vice President, Indiana University, Bloomington [d] Department of Communicative Sciences and Disorders, Michigan State University, East Lansing

ARTICLE INFO

Article History:
Received January 18, 2022
Revision received April 5, 2022
Accepted July 11, 2022

Editor-in-Chief: Peggy B. Nelson
Editor: Susan L. Thibeault

https://doi.org/10.1044/2022_JSLHR-22-00012

ABSTRACT

Purpose: Dysphonic voices typically present multiple voice quality dimensions. This study investigated potential interactions between perceived breathiness and roughness and their contributions to overall dysphonia severity.
Method: Synthetic stimuli based on four talkers were created to systematically map out potential interactions. For each talker, a stimulus matrix composed of 49 stimuli (seven breathiness steps × seven roughness steps) was created by varying aspiration noise and open quotient to manipulate breathiness and superimposing amplitude modulation of varying depths to simulate roughness. One-dimensional matching (1DMA) and magnitude estimation (1DME) tasks were used to measure perceived breathiness, roughness, their potential interactions, and overall dysphonia severity. Additional 1DME tasks were used to assess a set of natural stimuli that varied along both breathiness and roughness.
Results: For the synthetic stimuli, the 1DMA task indicated little interaction between the two voice qualities. For the 1DME task, breathiness magnitude was influenced by roughness step to a greater extent than roughness magnitude was influenced by breathiness step. The additive contributions of breathiness and roughness to overall severity gradually diminished with increasing breathiness and roughness steps, possibly reflecting a ceiling effect in the 1DME task. For the natural stimuli, little consistent interaction was observed between breathiness and roughness.
Conclusions: The matching task revealed minimal interaction between perceived breathiness and roughness, whereas the magnitude estimation task revealed some interaction between the two qualities and their cumulative contributions to overall dysphonia severity. Task differences are discussed in terms of differences in response bias and the role of perceptual anchors.
Supplemental Material: https://doi.org/10.23641/asha.21313701

Typical and dysphonic voices are characterized by changes along multiple voice quality (VQ) dimensions, each of which can vary relative to the other. VQ in dysphonic speakers is routinely evaluated in terms of three primary dimensions: breathiness, roughness, and strain (Barsties & De Bodt, 2015; Kempster et al., 2009; Shrivastav, 2011). Variations in these VQ dimensions can arise from a wide variety of pathologies associated with anatomical and physiological changes to the vocal folds or the vocal tract. Such changes can lead to complex perceptual changes that may differentially affect multiple VQ dimensions (Hirano, 1981; Holmberg et al., 2001; Morrison, 1997). In clinical practice, it is typical to use a single set of voice samples produced by a patient to make real-time perceptual judgments across individual VQ dimensions and judgments of the overall dysphonia severity. Although it is known that listeners can judge the magnitude of one auditory-perceptual dimension

Correspondence to Yeonggwang Park: park21@usf.edu. ***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

separately from another (e.g., pitch and loudness of the same sound; Cohen, 1961), whether and how changes across one VQ dimension affect the judgment of another dimension are not well understood. Since voice disorders rarely affect a single VQ dimension (e.g., some extremely rough voices may also be extremely breathy), it is important to understand how the magnitude of one VQ dimension may impact the perceptual judgments on another VQ dimension. A strong correlation among VQ dimensions reported in some previous studies (e.g., Pearson's r = .92 between auditory-perceptual ratings of breathiness and roughness; Ford Baldner et al., 2015) might indicate a common pathology impacting both VQ dimensions, one VQ dimension influencing the perceived magnitude of another dimension, or some combination of both.

The knowledge of such potential interactions between VQs is essential to fully understand their impact on auditory-perceptual and computational evaluation methods and downstream use of both methods in clinical assessments and research investigations. The overall goal of this study is to develop a more complete understanding of how perceived breathiness and roughness may co-occur, covary, and interact with each other in sets of well-controlled synthetic stimuli and natural voices that span wide ranges of perceived breathiness and roughness. When quantifying potential interactions between various VQ dimensions, it also is important to understand how those (putative) interactions relate to the scores listeners give to the overall severity of dysphonia.

While dysphonic VQ is commonly described along three dimensions (breathiness, roughness, and strain), the focus of this study was to understand the interaction between breathy and rough qualities only. These were selected due to the high prevalence and co-occurrence of these quality dimensions in the clinical population. Breathiness is defined as "audible air escape in the voice," and roughness is defined as "perceived irregularity in the voicing source" (Kempster et al., 2009). Organic or structural changes in the vocal folds as well as functional and neurological conditions that inhibit complete glottal adduction and lead to air escape through the vocal folds often result in increased aperiodic noise and breathiness (Carding et al., 2017; Hartl et al., 2001; Roy, 2008; Verdolini et al., 2006). Certain vocal pathologies such as nodules, polyps, and hemorrhage also impede symmetrical vocal fold vibration and can result in irregular vocal fold vibration and amplitude and frequency fluctuations in produced voice (Zhang et al., 2011), associated with perceived roughness (Awan & Awan, 2020; Latoszek, Maryn, et al., 2018). Because breathiness and roughness frequently coexist in disordered voices, a common term, "hoarseness," is thought to be a combination of the two qualities (Fairbanks, 1960; Kempster et al., 2009).

To best evaluate the interaction between breathiness and roughness, reliable and precise auditory-perceptual evaluation methods are necessary. We have previously demonstrated highly reliable judgments of breathy and rough VQ using a psychophysical matching task (e.g., Patel et al., 2010, 2012b). The matching tasks as implemented here quantify the magnitude of a VQ dimension by comparing the quality of a test stimulus against that of a speech-like comparison sound while systematically changing a single, independent parameter of the comparison sound. For breathy VQ, the independent variable used to match VQ perception is signal-to-noise ratio (SNR; Eddins et al., 2016; Patel et al., 2012a). This parameter was chosen based on the fact that SNR is analogous to harmonic-to-noise ratio, which often is computed as an acoustic correlate of breathiness (Hillenbrand, 1988; Kreiman & Gerratt, 2012), and because it has been shown to have high interrater and intrarater reliability and correlates well with other perceptual measures of breathiness (Patel et al., 2010). Similarly, for rough VQ, the independent variable used to match VQ perception is amplitude modulation (AM) depth, specifically at a modulation frequency of 25 Hz (Eddins & Shrivastav, 2013; Patel et al., 2012b) because of its strong relation to amplitude perturbations and prior work in sound quality that has used modulation depth successfully to quantify and model roughness of nonspeech sounds (e.g., Eddins & Shrivastav, 2013; Fastl & Zwicker, 2007; Patel et al., 2012b). Like SNR for breathiness, AM depth (in dB) was shown to have high interrater and intrarater reliability and to correlate well with other perceptual measures of roughness (Eddins & Shrivastav, 2013; Patel et al., 2012b). Furthermore, the relationship between perceived roughness and AM depth in synthetic stimuli, modeled after dysphonic voices, was strong (Eddins et al., 2015) and followed predictions consistent with the data of Eddins and Shrivastav (2013), Fastl and Zwicker (2007), and Patel et al. (2012b) regarding the perception of roughness in tonal stimuli.

For both quality dimensions, the value of the independent variable of the comparison sound at the point of subjective equality is taken as a measure of the magnitude of that VQ dimension with the specified physical units. Thus, breathiness is measured in units of SNR (in dB), whereas roughness is measured in units of AM depth (also quantified in dB). These matching tasks result in greater listener reliability and agreement for VQ dimensions of breathiness and roughness (intraclass correlation [ICC, 2, k] = 0.84–0.98) relative to the ordinal- and interval-level scales that are used in many clinical and research studies (e.g., Anand et al., 2019; Patel et al., 2010; Zraick et al., 2011). In addition to superior reliability, these one-dimensional matching (1DMA) tasks are by their nature dimension specific, are largely context independent and more robust to the order- or frequency-effects seen in rating data, and allow measurement of VQ magnitude using a ratio scale (Patel et al., 2012a, 2012b). Although 1DMA cannot be completed in real time with a patient in a clinical setting and takes longer than conventional rating scales, it can provide rigorous quantification of VQ in laboratory studies, on a ratio-level scale, with physical units that are meaningfully related to the underlying percept.

Another psychophysical task used to measure VQ is the unanchored one-dimensional magnitude estimation (1DME) task. In this task, listeners assign a number between 1 and 1,000 to indicate the perceived magnitude of a VQ dimension associated with a stimulus (Eddins et al., 2016; Patel et al., 2010; Shrivastav et al., 2011). While the 1DME also allows measurement along a ratio scale, this method requires assignment of arbitrary numbers to represent VQ magnitude and lacks a specific reference point. Such arbitrary assignment of numbers makes this method prone to certain biases caused by differences in individual internal standards and the contexts of experiments (Gerratt et al., 1993; Kreiman et al., 1992). While such arbitrary assignment of numerical values to represent VQ magnitude creates certain limitations, this task remains useful when matching may not be possible (e.g., when a suitable matching stimulus is unavailable or when one needs to study a large number of stimuli in a single, short-duration experiment). In this study, both of these psychophysical methods were adopted to investigate the interactions between VQ dimensions.

One challenge in studying the interaction between VQ dimensions is the identification of stimuli that vary systematically across just one or multiple VQ dimensions. This is further compounded by other changes in the voice stimuli, such as their fundamental frequency, formants, and spectral tilt. Each of these variables might further confound the results of an experiment designed to understand interaction across VQ dimensions. Such challenges may be addressed through the use of synthetically generated voices that allow precise control over all voice parameters and permit generation of stimuli that vary systematically along with one or multiple VQ dimensions. A second approach is to use natural voices, carefully selected from a large database using a stratified random sampling technique designed to identify a small set of stimuli that vary systematically across breathy and rough dimensions. Use of stratified sampling from existing databases has also been previously used to ensure that a wide range of VQ magnitudes is included in a set of test stimuli (e.g., Eadie & Baylor, 2006; Heman-Ackah et al., 2002; Hillenbrand & Houde, 1996; Shrivastav, 2003; Shrivastav & Sapienza, 2003; Shrivastav et al., 2005). While this approach helps select stimuli that vary systematically in VQ, it might not control for other variables (e.g., fundamental or formant frequencies). In this study, we used both approaches to investigate the potential covariance and interaction of breathy and rough VQs by modifying synthetic voice samples in a systematic manner and stratified sampling natural voice samples from databases of dysphonic voices. The synthetically generated stimuli allowed complete control of stimulus parameters—all stimuli were identical in all respects except for changes to the variable of interest. The natural stimuli selected through stratified-random sampling varied in more ways than just their VQ, but it was assumed as in many experiments on VQ that this variability had no or relatively small impact on listener judgments of breathiness and roughness.

Three experiments were conducted to investigate the interaction between breathiness and roughness. Experiments 1 and 2 used the same set of synthetic stimuli designed to allow precise control over the breathy and rough qualities while eliminating any other covariates (intensity, fundamental frequency, formants, etc.). For each quality dimension, the synthetic stimuli varied from low to high in equal physical steps. To measure perceived breathiness and roughness, Experiment 1 used two separate 1DMA tasks and Experiment 2 used two separate 1DME tasks. Because no matching paradigm has been developed for indexing overall severity, the 1DME task in Experiment 2 was included to capture overall severity and as a secondary method for evaluating potential interactions. We evaluated the null hypotheses that the presence of one VQ dimension in varying degrees of severity would not impact the perception of the other VQ dimension. We also evaluated the hypothesis that increases in both breathiness and roughness would contribute to overall severity. Finally, in Experiment 3, natural voice stimuli that varied in breathiness and roughness were used to evaluate whether the perception of the two VQs observed in Experiments 1 and 2 would extrapolate to natural stimuli.

Experiment 1: Evaluating Possible Interactions Between Perceived Breathiness and Roughness in Synthetic Stimuli Using 1DMA Tasks

Method

Listeners

Eight individuals (seven women, one man; age range: 19–31 years; mean age = 21.6 years) participated as "listeners" in this study.[1] All participants were native speakers of American English, had normal hearing (air-conduction pure-tone thresholds below 20-dB HL at 250, 500, 1000,

[1] Nine individuals were originally recruited, but one participant did not complete the entire experiment.

Table 1. Parameters used in the Klatt synthesizer to generate the four synthetic talkers.

Talker	Sex	f_o (Hz)	AV	SQ	F1 (B1) [Hz]	F2 (B2) [Hz]	F3 (B3) [Hz]	F4 (B4) [Hz]	AH range (step size)	OQ range (Step size)
003	F	223.3	60	250	976 (177)	1591 (150)	3107 (216)	4451 (574)	55–80 (4.2)	71–99 (4.7)
004	F	199.4	60	350	979 (250)	1436 (300)	2625 (581)	3546 (472)	50–80 (5)	66–99 (5.5)
007	M	117.6	60	400	741 (85)	1151 (140)	2485 (406)	3693 (536)	50–80 (5)	66–99 (5.5)
087	M	100.9	60	300	688 (84)	1022 (124)	2439 (202)	3329 (400)	55–80 (4.2)	71–99 (4.7)

Note. f_o = fundamental frequency; AV = amplitude of voicing; SQ = speed quotient; F1–F4 = formant frequencies; B1–B4, formant bandwidths; AH = amplitude of aspiration noise; OQ = open quotient; F = female; M = male.

2000, and 4000 Hz; American National Standards Institute [ANSI], 2010), and were students in the Department of Communicative Sciences and Disorders at Michigan State University.[2] This study was approved by the institutional review board, all participants consented to participate, and they were paid for their participation.

Synthetic Talker Stimuli

Four synthetic /ɑ/ vowels were modeled after four natural voices (two men, two women) from the University of Florida Disordered Voice Database as the sound source using a Klatt synthesizer with the Liljencrants-Fant model (Klatt & Klatt, 1990; Shrivastav & Camacho, 2010). Table 1 depicts parameters used in the Klatt synthesizer to generate the four synthetic talkers' voices. With these four synthetic vowels as the base stimulus, a set of stimuli varying in both breathiness and roughness was generated for each synthetic talker stimulus. Each set included seven discrete steps varying from low to high breathiness or low to high roughness. The breathiness set was generated by systematically increasing the amplitude of aspiration noise (AH) and the open quotient (OQ), which resulted in a lower SNR and a greater spectral slope, both of which are known to correlate with increasing breathiness (Hillenbrand et al., 1994; Klatt & Klatt, 1990; Shrivastav, 2003). The roughness set was generated by amplitude modulating the waveform of the synthetically generated vowels with systematically increasing depths of modulation. Increasing AM depth for relatively low-AM frequencies (25–50 Hz) correlates with the perception of roughness magnitude (Eddins et al., 2015; Fastl & Zwicker, 2007). Therefore, the vowel waveform was shaped by a complex AM function (a sinusoidal waveform raised to a power of 4 with a modulation frequency of 25 Hz), systematically increasing modulation depth from low to high. Based on our previous work (Eddins & Shrivastav, 2013), this manipulation resulted in a range of roughness that spans the range heard in normal and dysphonic voices as determined via a 1DMA task. Adjacent breathiness and roughness steps were equidistant from each other in physical units (i.e., SNR or AM depth) but these equal physical steps do not necessarily equate to equal perceptual steps. The stimulus matrix for each talker had 49 stimuli (seven steps of breathiness × seven steps of roughness), which resulted in a total of 196 stimuli (four synthetic talkers × 49 stimuli). All stimuli were cropped to include only the middle 500 ms and were shaped with a 10-ms cosine-squared window (see Supplemental Materials S1–S9 for stimuli examples).

Instrumentation

Stimulus presentation and response collection for all the auditory-perceptual tasks were controlled by the TDT SykofizX software application (Tucker-Davis Technologies, Inc.), which also controlled a TDT RZ6 real-time processor. The stimuli were delivered monaurally via high-fidelity insert earphones (ER2, Etymotic Research Inc.). The output level was calibrated to ensure that each stimulus was delivered at 85 dB SPL. All experimental procedures were conducted in a sound-attenuating booth, and the participants performed the 1DMA tasks using a computer monitor and a mouse inside the booth.

Procedure

All participants completed separate 1DMA tasks for breathiness and roughness over 14 to 24 test sessions depending on their speed of task performance. Each session spanning approximately 2 hr was scheduled over 2–2.5 months depending on participants' availability. The average total test duration for each participant in Experiment 1 was approximately 40 hr. The matching task required the comparison of a test stimulus with a

[2] At the time of data collection, the second and third authors were students and the fourth and fifth authors were faculty members at this institution.

synthetic comparison stimulus. The synthetic comparison sound was created using a low-pass-filtered sawtooth waveform (f_o = 151 Hz) added to a broadband speech-shaped noise that also was low-pass-filtered to have the same spectral envelope as the sawtooth wave. For breathiness, the single-variable parameter was the signal-to-noise ratio in decibels (dB SNR). For roughness, the single-variable parameter was the AM depth (in dB) of a 25-Hz modulator consisting of a sine function raised to a power of 4. The same background noise used as the independent variable for breathiness matching was added at a constant SNR of 20 dB for naturalness (Patel et al., 2012a, 2012b). On each trial, one of the 196 synthetic talker stimuli was presented, followed by 500 ms of silence and then the comparison stimulus, which was also 500 ms. The variable parameter of the comparison stimulus was increased or decreased (in 2-dB steps) with an up–down adaptive tracking procedure. The task was a modified two-alternative forced choice in which the independent variable of the comparison stimulus was either “increased” or “decreased” until a perceptual match was achieved, at which point the “equal” alternative was chosen. A total of six perceptual matches were obtained for each stimulus, three in which the initial independent variable value was at the high end of the continuum and three matches in which the initial independent variable value was at the low end of the continuum. The final match was taken as the average of the six matching values for each stimulus.

During each test session, the participants were trained to use the graphical user interface of the 1DMA task with synthetic sawtooth waveforms as the test stimuli to ensure that they could appropriately compare the variable parameters (introduced as signal noise for breathiness and fluctuation strength for roughness) between the test and comparison stimuli. They also performed practice 1DMA tasks with two voice samples that were not part of the experimental stimuli, before evaluating the experimental stimuli. The order of talkers, stimuli, and VQ dimensions were randomized across listeners. All data were collected for a single talker before proceeding to the next talker.

Statistical Analysis

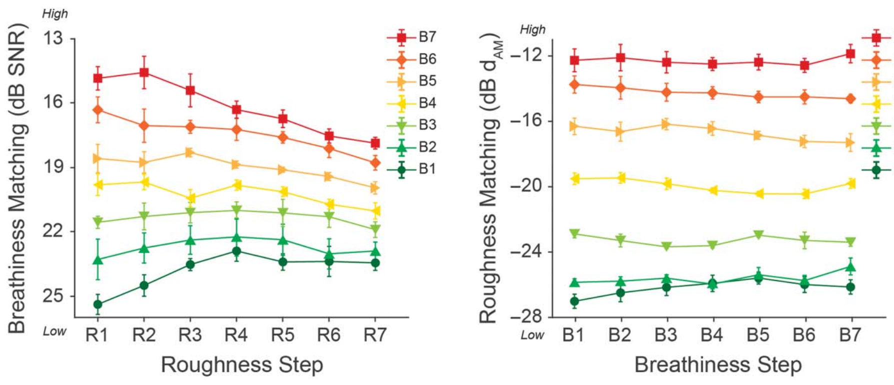
For each voice quality, ICC coefficients (2, *k*, absolute agreement) were used to calculate both intrarater and interrater reliability for the perceptual data (Shrout & Fleiss, 1979). Analysis of variance (ANOVA) was performed on mean breathiness and roughness matching values to determine the effects of talker, breathiness step, roughness step, and the interactions on perceived breathiness and roughness.

Results

Perceived Breathiness

For perceived breathiness, mean intralistener reliability, ICC(2, *k*), *k* = 6 reps, absolute agreement, was 0.96 and interlistener reliability, ICC(2, *k*), *k* = 8 listeners, absolute agreement, was 0.82. To examine the effect of roughness step on breathiness perception, mean breathiness matching values in dB SNR are plotted as a function of roughness step in the left panel of Figure 1. The averaged matching values across eight listeners and the four talkers are shown with standard error bars for each of the 49 breathiness–roughness step combinations. In this case, high perceived breathiness is associated with low SNR values, as shown toward the upper end of the *y*-axis. Low perceived breathiness is associated with high SNR values, as shown toward the lower end of the *y*-axis. For most of

Figure 1. Left panel: breathiness matching values (mean ± *SE* in dB SNR) as a function of roughness step (R1–R7) for each breathiness step (B1–B7). Right panel: roughness matching values (mean ± *SE* in dB modulation depth [dB d_{AM}]) as a function of breathiness step (B1–B7) for each roughness step (R1–R7).



the breathiness contours (i.e., B2, B3, B4, and B5), there appears to be little influence of roughness step on the breathiness matching values.

The effects of breathiness step, roughness step, and any possible interactions between breathiness step and roughness step were evaluated by computing a three-factor (talker, breathiness step, and roughness step) repeated-measures ANOVA. The results indicated that perceived breathiness differed significantly as a function of talker ($F_{3,22} = 12.01$, $p < .001$), breathiness step ($F_{6,18} = 116.78$, $p < .001$), and roughness step ($F_{6,18} = 8.32$, $p < .001$). A statistically significant interaction effect between breathiness and roughness steps was also observed ($F_{36,108} = 6.93$, $p < .001$). To explore the interaction between breathiness and roughness steps, seven post hoc one-way ANOVAs with roughness step as a factor were performed on mean breathiness matching values in each breathiness step (B1–B7), and the Benjamini–Hochberg procedure (Benjamini & Hochberg, 1995) was used to control Type I error. The effect of roughness step was only significant for the B7 samples ($p < .001$), whereas the effect of roughness step was not significant for samples B1–B6.

Perceived Roughness

Mean intralistener reliability, ICC(2, k), $k = 6$ reps, absolute agreement, was 0.99, and interreliability, ICC(2, k), $k = 8$ listeners, absolute agreement, was 0.97 for perceived roughness. To examine the effect of breathiness step on roughness perception, mean roughness matching values in dB modulation depth (dB d_{AM}) are plotted as a function of breathiness step in the right panel of Figure 1. The perception of roughness was minimally impacted by the co-occurrence of breathiness, as revealed by relatively flat contours of matching values across roughness steps.

To evaluate the effects of the breathiness step, roughness step, and any possible interactions, a three-factor (talker, roughness step, and breathiness step) repeated-measures ANOVA was computed. Roughness matching values differed significantly as a function of talker ($F_{3,21} = 10.25$, $p < .001$) and roughness step ($F_{6,18} = 644.11$, $p < .001$) but not as a function of breathiness step ($F_{6,18} = 0.74$, $p = .62$). There was a significant interaction between roughness and breathiness steps ($F_{36,108} = 2.24$, $p < .001$). To explore the interaction between breathiness and roughness steps, seven post hoc one-way ANOVAs with breathiness step as a factor were performed on mean roughness matching values for each roughness step (R1–R7). The effect of breathiness step was not significant for any roughness step, which indicates that roughness matching is not influenced by differing degrees of simulated breathiness from low to high.

Overall, these data indicate little interaction between breathiness step (systematic variations in AH and OQ yielding systematic differences in SNR) and roughness step (systematic variations in AM depth), with significant interactions observed for a minority (one of 14; B7) conditions.

Experiment 2: Evaluating Possible Interactions Between Perceived Breathiness and Roughness in Synthetic Stimuli Using 1DME Tasks

Method

Experiment 2 consisted of the same listeners, stimuli, and instrumentation as Experiment 1. Experiment 2 was completed over six to eight sessions (each < 2 hr) following the completion of Experiment 1 and differed only in the psychophysical procedure.

Procedure

Auditory-perceptual judgments were obtained using dimension-specific 1DME tasks. For each task, listeners estimated perceived breathiness, roughness, or overall dysphonia severity of each stimulus using a number between 1 and 1,000. Listeners were instructed that their estimate should reflect the ratio of VQ dimensions across samples. For example, a stimulus perceived to be twice as breathy as another stimulus would have to be given double the score. To familiarize the listeners with the magnitude estimation (ME) task, a loudness ME task involving nine tones varying in level from 60 to 92 dB SPL was performed before the actual 1DME tasks. For the primary tasks, each stimulus was presented 10 times in a random order (196 stimuli × 10 repetitions) and responses were averaged for each stimulus. Similar to Experiment 1, the order of talkers, stimuli, and VQ dimensions were randomized across listeners, and a stimulus set of a single talker was completed before proceeding to the next talker stimuli within a VQ dimension. Overall dysphonia severity was evaluated last after evaluation of either breathiness or roughness in random order, and the ME task on each VQ dimension was completed across two to three sessions within a week for most participants. The average duration for the three 1DME tasks was approximately 12 hr for each participant.

Statistical Analysis

All data were log-transformed (base 10) prior to statistical analysis. Furthermore, the data were normalized across listeners as follows in an effort to minimize the possible effects of different numerical ranges used by the listeners. First, the within-listener mean ME value for all 196 stimuli was computed. Second, the across-listener mean ME value was computed. Third, the difference between within-listener mean and across-listener mean was computed for each listener. Finally, for each listener, that difference was added to each of their 196 long-transformed judgments. Intralistener and interlistener reliability were measured using ICC (2, *k*, absolute agreement) for all perceptual data. ANOVA was used to determine the effects of talker, breathiness step, roughness step, and the interactions on perceived breathiness, roughness, and overall dysphonia severity. Additionally, a multiple regression was calculated to predict perceived overall dysphonia severity magnitudes based on breathiness and roughness steps.

Results

Perceived Breathiness

Mean intralistener reliability, ICC(2, *k*), k = 10 reps, absolute agreement, was 0.98, and interreliability, ICC(2, *k*), k = 8 listeners, absolute agreement, was 0.83 for perceived breathiness. To examine the effect of roughness step on breathiness perception, mean perceived breathiness magnitudes are plotted as a function of roughness step in the left panel of Figure 2. The breathiness magnitude judgments generally decreased gradually as the roughness step increased with the exception of the least breathy stimulus, B1, for which the breathiness magnitude judgments increased from R1 to R3 and decreased from R4 to R7.

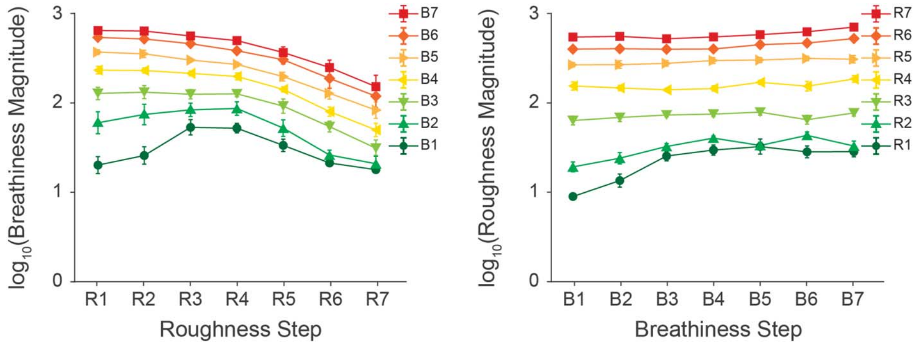
A three-way ANOVA (talkers, breathiness step, and roughness step as factors) indicated that perceived breathiness varied significantly among breathiness step ($F_{6,18}$ = 100.55, p < .001) and roughness steps ($F_{6,18}$ = 25.78, p < .001). There was also a significant interaction between breathiness step and roughness step ($F_{36,108}$ = 9.06, p < .001). To explore the interaction between breathiness and roughness steps, seven post hoc one-way ANOVAs with roughness step as a factor were performed on mean perceived breathiness magnitudes in each breathiness step (B1–B7). The effect of roughness step was significant for all breathiness functions (p < .001) and was mostly negative. However, for B1 samples, the perceived breathiness increased significantly when roughness step increased from R1 to R4, as shown in the bottom function in the left panel of Figure 2, and the perceived breathiness decreased significantly when roughness step increased from R5 to R7.

Perceived Roughness

Intralistener reliability, ICC(2, *k*), k = 10 reps, was 0.98, and interlistener reliability, ICC(2, *k*), k = 8 listeners, was 0.89 for perceived roughness. To examine the effect of breathiness step on roughness perception, mean perceived roughness magnitudes are plotted as a function of breathiness step in the right panel of Figure 2. Perceived roughness magnitude did not change substantially as the breathiness step increased from left to right within each roughness step.

A three-way ANOVA (talkers, breathiness step, and roughness step as factors) indicated that perceived roughness varied significantly among breathiness steps ($F_{6,18}$ = 22.49, p < .001) and roughness steps ($F_{6,18}$ = 523.66, p < .001). There was a significant interaction between roughness and breathiness steps ($F_{36,108}$ = 6.70, p < .001). To explore the interaction between breathiness and roughness steps, seven post hoc one-way ANOVAs with breathiness step as a factor were performed on mean perceived roughness magnitudes for each roughness step (R1–R7). The effect of breathiness step was statistically significant for R1 (p < .001), R2 (p < .001), R6 (p = .006), and R7 samples (p = .005). The effect of breathiness step on perceived roughness was mostly positive, but among the R6 and R7 functions, only the magnitudes for the B7 step

Figure 2. Left panel: log-transformed breathiness magnitudes (mean ± *SE*) as a function of roughness step (R1–R7) for each breathiness step (B1–B7). Right panel: log-transformed roughness magnitudes (mean ± *SE*) as a function of breathiness step (B1–B7) for each roughness step (R1–R7).



were statistically higher than for the B1–B4 steps. Thus, the interaction was not uniform across the breathiness dimension.

Overall, these data indicate an interaction between breathiness step (systematic variations in AH and OQ yielding systematic differences in SNR) and roughness step (systematic variations in AM depth) on both perceived breathiness and roughness magnitudes. Post hoc tests revealed that the effect of roughness step on perceived breathiness was mostly negative, whereas the effect of breathiness step on perceived roughness was mostly positive.

Perceived Overall Dysphonia Severity

Mean intralistener reliability, ICC(2, *k*), *k* = 10 reps, absolute agreement, was 0.97, and interlistener reliability, ICC(2, *k*), *k* = 8 listeners, absolute agreement, was 0.80 for perceived overall dysphonia severity. To examine the effects of breathiness and roughness steps on overall dysphonia severity, mean perceived severity magnitudes are plotted as a function of roughness step in the left panel of Figure 3 and a function of breathiness step in the right panel of Figure 3. The left and right panels of Figure 3 display the same data but are plotted with different *x*-axes. Overall dysphonia increased as roughness (left panel) and breathiness (right panel) steps increased. A three-way ANOVA (talkers, breathiness step, and roughness step as factors) indicated that perceived overall dysphonia severity varied significantly among talkers ($F_{3,25}$ = 3.56, p = .015), breathiness steps ($F_{6,18}$ = 171.51, p < .001), and roughness steps ($F_{6,18}$ = 305.51, p < .001). There was a significant interaction between breathiness and roughness steps ($F_{36,108}$ = 47.293, p < .001). As shown in Figure 3, unlike the dimension-specific matching and ME data, each dependent variable function shows a positive slope (i.e., post hoc linear regression indicated the slope to be significantly greater than zero), indicating that breathiness and roughness combine cumulatively, leading to greater overall severity than was attributed to one dimension alone. The fact that those slopes decreased as the secondary parameter increased (i.e., B1–B7 in the left panel; R1–R7 in the right panel) indicates that the two voice qualities have greater additivity when both are small and less additivity as their severity increases.

A multiple linear regression analysis revealed that breathiness and roughness steps were significant predictors of overall dysphonia severity ($F_{2,193}$ = 462.01, p < .001) with an R^2 of .83. Predicted overall dysphonia severity is equal to 1.378 + 0.077 (breathiness step) + 0.122 (roughness step). The log-transformed magnitudes of overall dysphonia severity increased 0.077 for each breathiness step (p < .001) and 0.122 for each roughness step (p < .001). As shown in Figure 3, overall severity judgments increased slightly more steeply as roughness increased (left panel) than as breathiness increased (right panel). To further examine the contribution of each predictor to the overall severity, additional simple linear regressions of overall dysphonia severity indicated that roughness step accounted for a greater proportion of variance (R^2 = .59, p < .001) than breathiness step (R^2 = .24, p < .001).

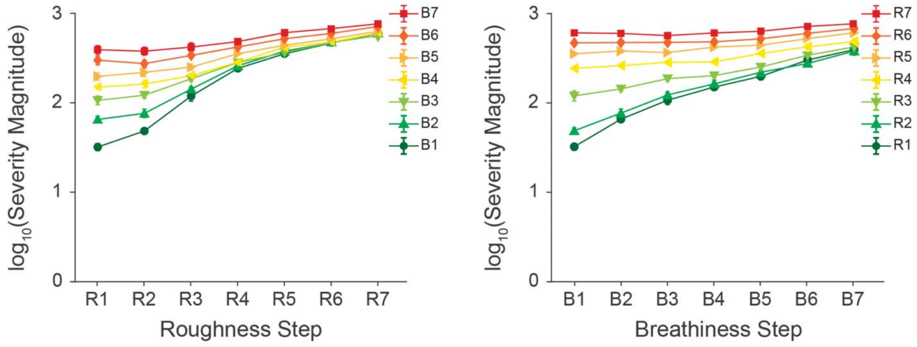
Experiment 3: Covariance in Natural Stimuli Using 1DME Tasks

Method

Listeners

Six women (mean age = 25.7 years) from the University of South Florida were recruited to participate in this study. All participants were native speakers of American English and had normal hearing (air-conduction puretone thresholds below 20 dB HL at 250, 500, 1000, 2000, and 4000 Hz; ANSI, 2010).

Figure 3. Left panel: log-transformed overall dysphonia severity magnitudes (mean ± *SE*) as a function of roughness step (R1–R7) for each breathiness step (B1–B7). Right panel: log-transformed overall dysphonia severity magnitudes (mean ± *SE*) as a function of breathiness step (B1–B7) for each roughness step (R1–R7). Both panels are displaying the same data but plotted with different *x*-axes.



Natural Stimuli

From three large disordered voice databases (University of Florida ENT database, Kay Elemetrics Disordered Voice database, and Sataloff/Heman-Ackah; Heman-Ackah et al., 2002), a total of 145 speakers' voices were identified through stratified random sampling by one of the authors (S.A.) such that dysphonic voices were primarily breathy and rough, with minimal other voice qualities such as strain, and represented a wide continuum of overall dysphonia severity. Following a consensus listening session by three of the authors (S.A., R.S., and D.A.E.), a two-dimensional matrix of 16 speakers' /ɑ/ recordings (four women, 12 men) was selected from the 145 voices such that each VQ dimension (breathiness and roughness) was sampled on a 4-point severity scale (none, mild, moderate, and severe). The speakers of the selected recordings had various voice pathologies including vocal hyperfunction, edema, unilateral and bilateral vocal fold paralysis, polyps, keratosis, presbylarynx, and laryngopharyngeal reflux.

Instrumentation

Stimulus presentation and response collection for the perceptual scaling tasks are identical to Experiments 1 and 2.

Procedure

Perceptual judgments of breathiness and roughness were obtained in three 1DME tasks: breathiness, roughness, and overall dysphonia severity. The experimental paradigm for the 1DME, including the loudness ME familiarization task, was identical to 1DME tasks with synthetic stimuli (Experiment 2). Each natural stimulus was presented 10 times in random order within a run (16 stimuli × 10 repetitions × 3 runs), and responses were averaged across 30 presentations per stimuli per listener. Data collection for each participant was completed over approximately 6 hr separated into three 2-hr sessions. All participants completed 1DME for overall dysphonia severity in the first session. The breathiness and roughness tasks were completed in random order across listeners in subsequent sessions.

Statistical Analysis

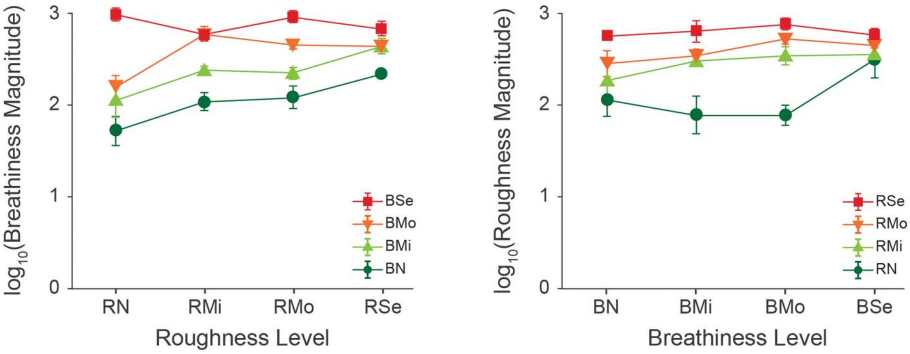
The same log transformation and normalization procedures applied in Experiment 2 were performed on the 1DME data. ICC (2, *k*, and absolute agreement) was used to calculate both intralistener and interlistener reliability for the perceptual data. ANOVA analyses were not completed due to having only one natural stimulus in each breathiness and roughness level. Instead, the trends in breathiness and roughness observed for the natural stimuli were compared against corresponding trends for synthetic stimuli (Experiment 2).

Results

Perceived Breathiness

Mean intralistener reliability, ICC(2, *k*), *k* = 30 presentations, absolute agreement, was 0.98, and interreliability ICC(2, *k*), *k* = 6 listeners, absolute agreement, was 0.98 for perceived breathiness. To examine the effect of roughness step on perceived breathiness, mean perceived breathiness magnitudes are plotted as a function of roughness step in the left panel of Figure 4. As shown in the breathiness contours (left panel), the positive slope indicates that samples with higher roughness were judged to have greater perceived breathiness relative to the samples with lower roughness for normal, mild, and moderate

Figure 4. Left panel: log-transformed breathiness magnitudes (mean ± *SE*) as a function of roughness level (RN–RSe) for each breathiness level (BN–BSe). Right panel: log-transformed roughness magnitudes (mean ± *SE*) as a function of breathiness level (BN–BSe) for each roughness step (RN–RSe). B = breathiness; R = roughness; N = none; Mi = mild; Mo = moderate; Se = severe.



breathiness samples. For the highest breathiness samples (BSe), variations in roughness did not alter breathiness magnitude estimates in a uniform manner.

Perceived Roughness

Mean intralistener reliability, ICC(2, k), k = 30 presentations, absolute agreement, was 0.97, and interreliability, ICC(2, k), k = 6 listeners, absolute agreement, was 0.91 for perceived roughness. To examine the effect of breathiness step on perceived roughness, mean perceived roughness magnitudes are plotted as a function of breathiness step in the right panel of Figure 4. As shown in the roughness contours, perceived roughness magnitude was minimally impacted by variation in breathiness magnitude from BN to BSe at least for the three higher roughness severities.

Perceived Overall Dysphonia Severity

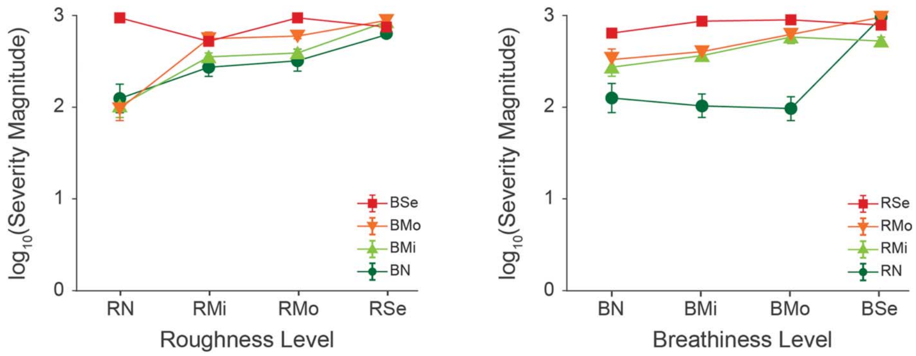
Mean intralistener reliability, ICC(2, k), k = 30 presentations, absolute agreement, was 0.97, and interreliability, ICC(2, k), k = 6 listeners, absolute agreement, was 0.97 for perceived dysphonia severity. To examine the effects of breathiness and roughness steps on perceived overall dysphonia severity, mean perceived severity magnitudes are plotted as a function of roughness step in the left panel of Figure 5 and as a function of breathiness step in the right panel of Figure 5. Data in the left panel reveal that, with increased roughness, perceived dysphonia severity increased for the three lower breathiness levels but not for the most severe degree of breathiness. The data are similar in form to dimension-specific 1DME (i.e., see Figure 4); increased roughness resulted in greater increases in overall severity than increased breathiness. Overall, the data reveal joint contributions of breathiness and roughness levels to overall dysphonia severity judgments. The relatively sparse sampling on both dimensions and irregular patterns of change in magnitude, however, make further characterization difficult. Furthermore, the fact that the highest levels in each dimension resulted in values near the maximum allowable magnitude estimate (1000; $\log_{10}[1000] = 3$) indicates a ceiling effect that might have limited the ability to observe variation along the complimentary dimension.

Discussion

This study evaluated potential interactions between perceived breathiness and roughness in sets of synthetic and natural voice stimuli using two different psychophysical tasks, matching and ME. For the synthetic stimuli, breathiness and roughness steps were obtained for each of four talkers by creating seven linearly spaced stimuli: manipulating OQ and aspiration noise to result in seven breathiness steps and manipulating AM depth to result in seven roughness steps. Though not evaluated here, there was no assumption that the equal physical step sizes would result in perceptually equal steps. As expected, differences among the seven synthetic stimuli in breathiness or roughness continuum were easily identified through informal listening and as revealed in the results of two different psychophysical tasks. A smaller set of natural voices was included in Experiment 3 as a means of cross-validation of the results using synthetic stimuli.

Most previous investigations using a single-variable matching task have used natural stimuli divided into sets of primarily breathy or primarily rough stimuli spanning a wide range of dysphonia from minimal to severe. By design, those investigations could not evaluate the possibility of interaction among voice qualities. In this investigation, the

Figure 5. Left panel: log-transformed overall dysphonia severity magnitudes (mean ± *SE*) as a function of roughness level (RN–RSe) for each breathiness level (BN–BSe). Right panel: log-transformed overall dysphonia severity magnitudes (mean ± *SE*) as a function of breathiness level (BN–BSe) for each roughness step (RN–RSe). Both panels are displaying the same data but plotted with different *x*-axes. B = breathiness; R = roughness; N = none; Mi = mild; Mo = moderate; Se = severe.



synthetic stimuli were specifically generated to evaluate the interaction between breathiness and roughness perception. The resulting matching data show minimal interaction between the two VQ dimensions. The matching task revealed only one stimulus series—the one with the highest magnitude of breathiness—for which there was a statistically significant impact of roughness step on the breathiness matching magnitude. In contrast, variations in roughness had no statistically significant impact on the matching magnitude of breathiness. The results might be explained based on the acoustic characteristics manipulated in this experiment. Aspiration noise is greatest in the mid-to-high frequency range and might not markedly impact perceived roughness, which often is associated with low-frequency aperiodic noise (Latoszek, Maryn, et al., 2018). OQ alters the spectral slope of the stimulus, a property not strongly associated with perceived roughness (Latoszek, de Bodt, et al., 2018).

The 1DME tasks in Experiment 2 revealed an inverse effect of roughness magnitude on the estimation of breathiness magnitude (see Figure 2, left panel) and some impact of breathiness on roughness magnitude. While not dramatic, the differences between the 1DME and the 1DMA data are intriguing and deserve consideration. From a task perspective, the ME task is more susceptible to set biases than the matching task. Due to the vast number of stimuli, they were blocked by talker such that all ME judgments were completed for one talker before proceeding to the next talker. It is possible that imposition of this restriction in the design might have accentuated potential biases and influenced 1DME judgments in a way that was not revealed in 1DMA judgments. Similarly, while a statistical correction was completed to normalize the range of numbers assigned by different listeners and in different experimental sessions, this normalization cannot correct for all biases.

The nature of the synthetic stimuli also may have differentially impacted the results of the 1DME and the 1DMA tasks. When synthesizing the roughness stimulus series, AM was applied after vowel synthesis, and thus, it was superimposed equally on the aspiration noise, as well as the harmonic components in the stimuli. During natural voicing, however, this may not be the case. Since aspiration noise often arises from the DC flow through the glottal gap, it is possible that the aspiration noise is not modulated in the same way as the vowel harmonics. It is possible that such subtle changes impact the naturalness of the stimuli and also the perception of roughness or breathiness itself. Such equal superposition of AM on both the periodic and aperiodic components of the stimuli might have artificially introduced the interaction that was observed for the highest breathiness step from the 1DMA task. Because comparison stimuli used in the roughness matching task were synthesized in a manner similar to synthesis of the roughness steps of the synthetic stimuli (i.e., application of AM to both the sawtooth and noise components), the common synthesis approach for the comparison and target sounds might have canceled out any perceptual impacts of applying modulation in equal proportions to the periodic and aperiodic components. Judgments in the 1DME task may have been sensitive to such synthesis parameters, which would be consistent with different degrees of interaction associated with the 1DMA and 1DME tasks.

Additionally, perceived breathiness may affect perceived roughness in some cases by decreasing the periodicity of the signal, as observed in some of the samples from the ME task. Unlike the perceived roughness result of the matching task, the result of the ME task showed that breathiness step statistically increased perceived roughness in R1, R2, R6, and R7 samples. Added aspiration noise reduces the periodicity of the signal, which may have contributed to increases in perceived roughness. The loss of periodicity was shown to have a statistically significant relationship to perceived roughness in conventional rating tasks although the correlation was weak to moderate (Bhuta et al., 2004; Latoszek, de Bodt, et al., 2018). Similarly, aspiration noise in natural disordered voices may also decrease the periodicity of the voices, and thus, the presence of breathiness in natural voices might augment perception of co-occurring roughness in clinical evaluation of VQ. However, this interaction seems minimal in matching tasks, as observed in Experiment 1, indicating that comparison sounds may have reduced the impact of co-occurring breathiness on roughness evaluation.

As noted previously, the matching task offers several advantages relative to other perceptual tasks when estimating the magnitude of a specific percept. Individual internal standards and biases have been suggested as factors that reduce the reliability of several other traditional auditory-perceptual tasks (Kreiman et al., 1992; Oates, 2009; Shrivastav et al., 2005). The matching task involves sequential presentation of the comparison stimulus and target stimulus on each trial, in which case the target sound functions as a trial-by-trial acoustic anchor. Previous studies have observed that auditory-perceptual tasks with samples that were available for listeners to compare have higher reliability than tasks without comparison samples (dos Santos et al., 2019; Kapsner-Smith et al., 2021; Kreiman & Gerratt, 2011). In addition, matching tasks allow modification of specific acoustic features that closely correspond to the VQ of interest and those features can then serve as an independent measure of the percept. Those acoustic modifications, or independent variables, have physical units associated with them that allow quantitative and meaningful capture of the perceptual attribute(s) being matched. These features of the matching task allow direct comparison of data obtained across experiments, stimuli, test sessions, or other intervening variables. However, compared to ME tasks or conventional ratings, the current matching task cannot provide a real-time measurement of VQ for clinical use due to the time required for measurement.

The matching task used in the current investigation, similar to previous studies, considered only one VQ dimension at a time (i.e., breathiness and roughness were measured in separate sessions on separate days). An alternative approach would be to use a two-dimensional matching task (2DMA) in which matching is based on two independent variables (i.e., dB SNR and dB modulation depth) that are manipulated either simultaneously in a single trial or sequentially in adjacent trials. The temporal proximity of breathiness and roughness judgments might provide a more accurate measurement of the perception of the two qualities within a single stimulus.

Experiment 3 tested the interaction between breathiness and roughness perception using natural voice samples and a 1DME task. This experiment did not show the inverse relationship between roughness and breathiness magnitude that was observed for the synthetic stimuli in Experiment 2. This observation further supports that the observed inverse relationship might be due to the nature of synthetic stimuli generation as described previously. While the smaller number of stimuli limits a more comparable statistical analysis to Experiments 1 or 2, the general patterns reveal little consistent interaction between breathiness and roughness perception.

Overall severity of dysphonia was considered in Experiments 2 and 3 using the 1DME task. There is no comparable matching task for assessment of overall dysphonia severity. The results of the 1DME task indicate that both breathiness and roughness contribute to perceptual judgments of overall severity and the relationship is additive in nature. This is consistent with previous work that also reported significant correlations between breathiness, roughness, and overall quality (Ford Baldner et al., 2015), indicating a contribution from each individual VQ to the overall dysphonia severity of voice. Variation in aspiration noise and OQ and AM depth resulted in significant increases in the perceived severity of dysphonia for all synthetic stimulus series tested in Experiment 2. However, the change in overall severity due to increases in breathiness or roughness step is smaller when the step value of the other dimension is already very high. Similar trends for overall dysphonia severity also were observed with the natural stimuli in Experiment 3; overall severity mostly increased at the normal, mild, and moderate levels of breathiness or roughness and did not change markedly when stimuli were at the severe level of breathiness or roughness. Thus, the two qualities are additive but severity judgments do not scale up indefinitely. Instead, the severity judgments show the greatest growth at low levels of roughness or breathiness and less growth for high levels of roughness or breathiness. Such compressive relationships also are seen for other psychophysical continua (e.g., the combined impact of frequency and sound pressure level on equal loudness contours; Fletcher & Munson, 1933). The compressive function observed here (e.g., see Figures 3 and 5) not only might reflect the nature of perceptual mechanisms but also might reflect limitations of the 1DME task such as ceiling effects. Using an unbounded ME task would overcome the latter limitation.

One limitation of this study is unbalanced distribution of sex among our listeners. Our listeners were mostly females, and in Experiments 1 and 2, there was one male listener. To the best of our knowledge, the effect of listener's sex on perception of VQ has not been examined; hence, it is unknown whether the sex imbalance influenced our results. We confirmed that the results of the male listener were within the range of the seven female listeners. Another potential limitation of the study is potential perceptual drift of participants over the long duration of Experiments 1 and 2. For matching tasks in Experiment 1, the listeners came in for 20 sessions on average across 2.5 months. However, the matching tasks provided comparison sounds as a reference, and thus, we expect that the matching results would not have been affected by internal perceptual standards drifted across the experimental sessions. We also aimed to minimize the possible effect of perceptual drift in Experiment 2 by scheduling participants' sessions to complete one VQ dimension within a week.

Conclusions

Using a 1DMA task, the data from this study indicate little interaction among the perceived breathiness and perceived roughness for synthetic stimuli designed to vary from minimal to extremely breathy or rough. Data from the 1DME tasks, however, indicate potential interactions among the two voice qualities in synthetic stimuli might affect the estimated magnitudes of the other VQ. Data from the 1DME tasks also indicated that increasing degrees of breathiness and roughness combined to contribute to progressive increases in overall severity of dysphonia in both synthetic and natural stimuli. Knowledge of the interactions between various VQ dimensions and their contributions to dysphonia severity is necessary to improve theoretical and computational models of the perception of dysphonia and for the development of clinical tools that more accurately index individual VQ dimensions.

Data Availability Statement

The published data are available from the corresponding author upon reasonable request.

Acknowledgments

This work was supported by the National Institute for Deafness and Other Communicative Disorders R01DC009029 (David A. Eddins and Rahul Shrivastav). The authors would like to thank Alissa Hebert, Meaghan Kilbourn, and Mckenna Dyjak for recruitment and data collection.

References

American National Standards Institute. (2010). *Methods for manual pure-tone threshold audiometry*.

Anand, S., Skowronski, M. D., Shrivastav, R., & Eddins, D. A. (2019). Perceptual and quantitative assessment of dysphonia across vowel categories. *Journal of Voice, 33*(4), 473–481. https://doi.org/10.1016/j.jvoice.2017.12.018

Awan, S. N., & Awan, J. A. (2020). A two-stage cepstral analysis procedure for the classification of rough voices. *Journal of Voice, 34*(1), 9–19. https://doi.org/10.1016/j.jvoice.2018.07.003

Barsties, B., & De Bodt, M. (2015). Assessment of voice quality: Current state-of-the-art. *Auris Nasus Larynx, 42*(3), 183–188. https://doi.org/10.1016/j.anl.2014.11.001

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bhuta, T., Patrick, L., & Garnett, J. D. (2004). Perceptual evaluation of voice quality and its correlation with acoustic measurements. *Journal of Voice, 18*(3), 299–304. https://doi.org/10.1016/j.jvoice.2003.12.004

Carding, P., Bos-Clark, M., Fu, S., Gillivan-Murphy, P., Jones, S. M., & Walton, C. (2017). Evaluating the efficacy of voice therapy for functional, organic and neurological voice disorders. *Clinical Otolaryngology, 42*(2), 201–217. https://doi.org/10.1111/coa.12765

Cohen, A. (1961). Further Investigation of effects of intensity upon pitch of pure tones. *The Journal of the Acoustical Society of America, 33*(10), 1363–1376. https://doi.org/10.1121/1.1908441

dos Santos, P. C. M., Vieira, M. N., Sansao, J. P. H., & Gama, A. C. C. (2019). Effect of auditory-perceptual training with natural voice anchors on vocal quality evaluation. *Journal of Voice, 33*(2), 220–225. https://doi.org/10.1016/j.jvoice.2017.10.020

Eadie, T. L., & Baylor, C. R. (2006). The effect of perceptual training on inexperienced listeners' judgments of dysphonic voice. *Journal of Voice, 20*(4), 527–544. https://doi.org/10.1016/j.jvoice.2005.08.007

Eddins, D. A., Anand, S., Camacho, A., & Shrivastav, R. (2016). Modeling of breathy voice quality using pitch-strength estimates. *Journal of Voice, 30*(6), 774.e1–774.e7. https://doi.org/10.1016/j.jvoice.2015.11.016

Eddins, D. A., Kopf, L. M., & Shrivastav, R. (2015). The psychophysics of roughness applied to dysphonic voice. *The Journal of the Acoustical Society of America, 138*(6), 3820–3825. https://doi.org/10.1121/1.4937753

Eddins, D. A., & Shrivastav, R. (2013). Psychometric properties associated with perceived vocal roughness using a matching task. *The Journal of the Acoustical Society of America, 134*(4), EL294–EL300. https://doi.org/10.1121/1.4819183

Fairbanks, G. (1960). *Voice and articulation drillbook* (2nd ed.). Harper.

Fastl, H., & Zwicker, E. (2007). *Psychoacoustics: Facts and models* (3rd. ed.). Springer. http://www.loc.gov/catdir/toc/fy0709/2006934622.html

Fletcher, H., & Munson, W. A. (1933). Loudness, its definition, measurement and calculation. *The Journal of the Acoustical Society of America, 5*(2), 82–108. https://doi.org/10.1121/1.1915637

Ford Baldner, E., Doll, E., & van Mersbergen, M. R. (2015). A review of measures of vocal effort with a preliminary study on the establishment of a vocal effort measure. *Journal of Voice, 29*(5), 530–541. https://doi.org/10.1016/j.jvoice.2014.08.017

Gerratt, B. R., Kreiman, J., Antonanzas-Barroso, N., & Berke, G. S. (1993). Comparing internal and external standards in voice quality judgments. *Journal of Speech and Hearing Research, 36*(1), 14–20. https://doi.org/10.1044/jshr.3601.14

Hartl, D. M., Hans, S., Vaissiere, J., Riquet, M., & Brasnu, D. F. (2001). Objective voice quality analysis before and after onset of unilateral vocal fold paralysis. *Journal of Voice, 15*(3), 351–361. https://doi.org/10.1016/S0892-1997(01)00037-6

Heman-Ackah, Y. D., Michael, D. D., & Goding, G. S., Jr. (2002). The relationship between cepstral peak prominence and selected parameters of dysphonia. *Journal of Voice, 16*(1), 20–27. https://doi.org/10.1016/s0892-1997(02)00067-x

Hillenbrand, J. (1988). Perception of aperiodicities in synthetically generated voices. *The Journal of the Acoustical Society of America, 83*(6), 2361–2371. https://doi.org/10.1121/1.396367

Hillenbrand, J., Cleveland, R. A., & Erickson, R. L. (1994). Acoustic correlates of breathy vocal quality. *Journal of Speech and Hearing Research, 37*(4), 769–778. https://doi.org/10.1044/jshr.3704.769

Hillenbrand, J., & Houde, R. A. (1996). Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech. *Journal of Speech and Hearing Research, 39*(2), 311–321. https://doi.org/10.1044/jshr.3902.311

Hirano, M. (1981). *Clinical examination of voice*. Springer-Verlag.

Holmberg, E. B., Hillman, R. E., Hammarberg, B., Sodersten, M., & Doyle, P. (2001). Efficacy of a behaviorally based voice therapy protocol for vocal nodules. *Journal of Voice, 15*(3), 395–412. https://doi.org/10.1016/S0892-1997(01)00041-8

Kapsner-Smith, M. R., Opuszynski, A., Stepp, C. E., & Eadie, T. L. (2021). The effect of visual sort and rate versus visual analog scales on the reliability of judgments of dysphonia. *Journal of Speech, Language, and Hearing Research, 64*(5), 1571–1580. https://doi.org/10.1044/2021_JSLHR-20-00623

Kempster, G. B., Gerratt, B. R., Verdolini Abbott, K., Barkmeier-Kraemer, J., & Hillman, R. E. (2009). Consensus auditory-perceptual evaluation of voice: Development of a standardized clinical protocol. *American Journal of Speech-Language Pathology, 18*(2), 124–132. https://doi.org/10.1044/1058-0360(2008/08-0017)

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *The Journal of the Acoustical Society of America, 87*(2), 820–857. https://doi.org/10.1121/1.398894

Kreiman, J., & Gerratt, B. R. (2011). Comparing two methods for reducing variability in voice quality measurements. *Journal of Speech, Language, and Hearing Research, 54*(3), 803–812. https://doi.org/10.1044/1092-4388(2010/10-0083)

Kreiman, J., & Gerratt, B. R. (2012). Perceptual interaction of the harmonic source and noise in voice. *The Journal of the Acoustical Society of America, 131*(1), 492–500. https://doi.org/10.1121/1.3665997

Kreiman, J., Gerratt, B. R., Precoda, K., & Berke, G. S. (1992). Individual differences in voice quality perception. *Journal of Speech and Hearing Research, 35*(3), 512–520. https://doi.org/10.1044/jshr.3503.512

Latoszek, B. B. V., de Bodt, M., Gerrits, E., & Maryn, Y. (2018). The exploration of an objective model for roughness with several acoustic markers. *Journal of Voice, 32*(2), 149–161. https://doi.org/10.1016/j.jvoice.2017.04.017

Latoszek, B. B. V., Maryn, Y., Gerrits, E., & de Bodt, M. (2018). A meta-analysis: Acoustic measurement of roughness and breathiness. *Journal of Speech, Language, and Hearing Research, 61*(2), 298–323. https://doi.org/10.1044/2017_JSLHR-S-16-0188

Morrison, M. (1997). Pattern recognition in muscle misuse voice disorders: How I do it. *Journal of Voice, 11*(1), 108–114. https://doi.org/10.1016/s0892-1997(97)80031-8

Oates, J. (2009). Auditory-perceptual evaluation of disordered voice quality: Pros, cons and future directions. *Folia Phoniatrica et Logopaedica, 61*(1), 49–56. https://doi.org/10.1159/000200768

Patel, S., Shrivastav, R., & Eddins, D. A. (2010). Perceptual distances of breathy voice quality: A comparison of psychophysical methods. *Journal of Voice, 24*(2), 168–177. https://doi.org/10.1016/j.jvoice.2008.08.002

Patel, S., Shrivastav, R., & Eddins, D. A. (2012a). Developing a single comparison stimulus for matching breathy voice quality. *Journal of Speech, Language, and Hearing Research, 55*(2), 639–647. https://doi.org/10.1044/1092-4388(2011/10-0337)

Patel, S., Shrivastav, R., & Eddins, D. A. (2012b). Identifying a comparison for matching rough voice quality. *Journal of Speech, Language, and Hearing Research, 55*(5), 1407–1422. https://doi.org/10.1044/1092-4388(2012/11-0160)

Roy, N. (2008). Assessment and treatment of musculoskeletal tension in hyperfunctional voice disorders. *International Journal of Speech-Language Pathology, 10*(4), 195–209. https://doi.org/10.1080/17549500701885577

Shrivastav, R. (2003). The use of an auditory model in predicting perceptual ratings of breathy voice quality. *Journal of Voice, 17*(4), 502–512. https://doi.org/10.1067/s0892-1997(03)00077-8

Shrivastav, R. (2011). Evaluating voice quality. In E. P. M. Ma & E. M. L. Yiu (Eds.), *Handbook of voice assessments* (pp. 305–318). Singular.

Shrivastav, R., & Camacho, A. (2010). A computational model to predict changes in breathiness resulting from variations in aspiration noise level. *Journal of Voice, 24*(4), 395–405. https://doi.org/10.1016/j.jvoice.2008.12.001

Shrivastav, R., Camacho, A., Patel, S., & Eddins, D. A. (2011). A model for the prediction of breathiness in vowels. *The Journal of the Acoustical Society of America, 129*(3), 1605–1615. https://doi.org/10.1121/1.3543993

Shrivastav, R., & Sapienza, C. M. (2003). Objective measures of breathy voice quality obtained using an auditory model. *The Journal of the Acoustical Society of America, 114*(4), 2217–2224. https://doi.org/10.1121/1.1605414

Shrivastav, R., Sapienza, C. M., & Nandur, V. (2005). Application of psychometric theory to the measurement of voice quality using rating scales. *Journal of Speech, Language, and Hearing Research, 48*(2), 323–335. https://doi.org/10.1044/1092-4388(2005/022)

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428. https://doi.org/10.1037//0033-2909.86.2.420

Verdolini, K., Rosen, C. A., & Branski, R. C. (2006). *Classification manual for voice disorders-I*. Erlbaum.

Zhang, Y., Shao, J., Krausert, C. R., Zhang, S., & Jiang, J. J. (2011). High-speed image analysis reveals chaotic vibratory behaviors of pathological vocal folds. *Chaos, Solitons & Fractals, 44*(1–3), 169–177. https://doi.org/10.1016/j.chaos.2011.01.007

Zraick, R. I., Kempster, G. B., Connor, N. P., Thibeault, S., Klaben, B. K., Bursac, Z., Thrush, C. R., & Glaze, L. E. (2011). Establishing validity of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V). *American Journal of Speech-Language Pathology, 20*(1), 14–22. https://doi.org/10.1044/1058-0360(2010/09-0105)