ASHA American Speech-Language-Hearing Association

JSLHR

Research Note

Evaluating the Effect of Voice Quality Covariance on Auditory-Perceptual Evaluation Using a Novel Two-Dimensional Magnitude Estimation Task

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

[a] Department of Communication Sciences & Disorders, University of South Florida, Tampa [b] Department of Communication Sciences and Disorders, University of Central Florida, Orlando [c] Office of the Provost & Executive Vice President, Indiana University Bloomington

ARTICLE INFO

Article History:
Received April 4, 2023
Revision received August 12, 2023
Accepted September 3, 2023

Editor: Susan L. Thibeault

https://doi.org/10.1044/2023_JSLHR-23-00226

ABSTRACT

Purpose: Most people with dysphonia present with voices that vary along more than one voice quality (VQ) dimension. This study sought to examine the effect of covariance between breathy and rough VQ in natural voices.
Method: A two-dimensional matrix of 16 /a/ vowels was selected such that two VQ dimensions (breathiness and roughness) were sampled on a 4-point severity scale (none, mild, moderate, and severe). Ten listeners evaluated 480 stimuli (16 stimuli × 10 repetitions × 3 blocks) on one-dimensional magnitude estimation (1DME) tasks and a novel two-dimensional magnitude estimation (2DME) task that allowed for simultaneous measurement of breathiness and roughness.
Results: Data indicated high intra- and interrater reliabilities for both breathiness and roughness in the 2DME and 1DME tasks. Correlation analyses revealed a strong correlation between 2DME and 1DME judgments for breathiness and roughness ($r > .95$). There was also a minimal correlation between breathy and rough VQ in the 2DME task ($r < .10$).
Conclusions: Covarying roughness or breathiness had less impact on the perception of the other VQ in natural dysphonic voices in 2DME compared to 1DME. An understanding and quantification of the perceptual interactions among the dimensions will aid in the refinement of computational models and in the establishment of the validity of clinical scales for VQ perception.

Voice quality (VQ) is a multidimensional perceptual construct, and its three primary dimensions are breathiness, roughness, and strain (American Speech-Language-Hearing Association, 2002; Hirano, 1981). Among these dimensions, breathiness and roughness are more commonly observed (Dejonckere, 1995), and often co-occur in many dysphonic voices. Therefore, this study focused on these two quality dimensions and their auditory-perceptual judgments. Breathy voices result from air leakage during glottal closure and are characterized by turbulent high-frequency noise during phonation. *Breathiness* has been defined as "audible air escape in the voice" (Kempster et al., 2009). Rough voices result from recurrent, rapid, and random changes to the habitual movement patterns of the vocal fold and are typically characterized as having low-frequency aperiodic noise with subharmonics. *Roughness* has been defined as "perceived irregularity in the voicing source" (Kempster et al., 2009). Although most pathological voices are characterized by the co-occurrence of breathiness and roughness, the interaction between these VQ dimensions is not well understood. For example, vocal fold paralysis may inhibit compete closure of the vocal folds, leading to excessive air escape and resulting in a breathy voice. Vocal fold paralysis may also affect the regularity of vibratory cycles, resulting in a rough voice. Given that several vocal pathologies may result in dysphonia affecting multiple VQ dimensions simultaneously, a systematic investigation and explanation of the covariance among VQ dimensions and their potential interactions is imperative for advancing voice research and clinical practice.

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

Auditory-perceptual judgments are a fundamental component of VQ measurement for clinical and research purposes. Current scales such as the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V; Kempster et al., 2009) or GRBAS (grade, roughness, breathiness, asthenia, strain; Hirano, 1981) used for the evaluation of VQ require clinicians to elicit voice samples such as sustained vowels, sentences, and a brief passage and then judge the voice in terms of parameters including overall severity, pitch, loudness, and the VQ dimensions of breathiness, roughness, and strain. To our knowledge, only two published studies have reported covariance among VQ dimensions and revealed moderate to strong correlations between breathy and rough VQ (Baldner et al., 2015; Walden & Rau, 2022). Baldner et al. (2015) obtained auditory-perceptual ratings (CAPE-V), vocal effort ratings (Borg CR-10), Voice Handicap Index, and phonation threshold pressure measures for 28 people with voice disorders and 28 healthy controls. Their results showed that the perceived breathiness and roughness were highly and significantly correlated (Pearson's $r = .92$, $p < .001$). Walden and Rau (2022) studied the contributions of individual CAPE-V and GRBAS parameters to overall dysphonia severity in 296 voice samples. Breathiness and roughness were moderately but significantly correlated (CAPE-V: Pearson's $r = .64$, $p < .01$; GRBAS: Pearson's $r = .56$, $p < .01$). Although these studies provide evidence and support that many dysphonic voices covary along multiple VQ dimensions, they did not examine or explain whether variability along one VQ dimension may have impacted the perceived magnitude of another dimension. For example, in some severely rough voices that are also breathy, would the roughness severity lead to increased severity rating of breathiness?

The covariance of multiple VQ dimensions that often occurs in natural dysphonic voices can be captured using basic psychophysical methods. One such classical method for measuring sensations and evaluating prothetic continua, such as VQ, which has been used extensively, is the magnitude estimation (ME) task (e.g., Eddins et al., 2016, 2021; Gescheider, 1976; Mckenna, 1985; Shrivastav & Camacho, 2010; Stevens, 1957, 1958). In one protocol for this task, listeners assign to each stimulus a number, often within a defined range such as 1 to 1,000, that indicates the perceived magnitude of the VQ under study. Listeners are instructed to choose numbers in a way that reflects the magnitude of the differences in VQ across stimuli. Experimenters ask listeners to assign numbers on a ratio scale, so that this task can provide ratio-level data compared to the ordinal or interval-level data from *N*-point Likert rating scales, GRBAS, and visual-analog scales in CAPE-V (Nagle, 2016; Patel et al., 2010). Since the procedure requires listeners to think in ratios or fractions instead of intervals as in visual analog scales such as the CAPE-V, the use of a large range from 1 to 1,000 is typically used. For example, evaluating the brightness of a light that is 20 times brighter than another would allow a listener to numerically assign a large number and accurately perceive the magnitude of the stimulus. Related methods include fractionation (Stevens, 1956, 1959) and magnitude production (Green et al., 1977; Stevens, 1957). Many studies in the field of VQ perception have used ME tasks to investigate one VQ dimension at a time (one-dimensional magnitude estimation [1DME], e.g., Hillenbrand et al., 1994; Shrivastav et al., 2011). This research has successfully shown that perceptual judgments of breathiness and roughness obtained using such a 1DME task are highly reliable within and across listeners (Eddins et al., 2021; Park et al., 2022). Furthermore, prior investigations on the properties of the ME task indicate that the results obtained depend very little on whether or not a predefined range is used and whether or not an anchor stimulus is used (e.g., Stevens, 1956).

In a systematic investigation of VQ covariance, Park et al. (2022) examined the interactions between breathy and rough voice dimensions and their overall contributions to dysphonia severity in synthetic stimuli ($N = 49$) and natural stimuli ($N = 16$) that were either primarily breathy or rough. Synthetic stimuli based on four talkers were created to generate a matrix of seven breathiness levels and seven roughness levels for each talker. Breathiness levels were created by manipulating the aspiration noise and open quotient, whereas roughness levels were created by manipulating the amplitude modulation depths. A 1DME task allowed listeners to evaluate each of the VQs in different sessions. For synthetic stimuli, the magnitude of breathiness was found to interact with the roughness level to a greater extent than the opposite. In other words, high levels of roughness impacted the magnitude estimates of breathiness, but high levels of breathiness did not affect the roughness magnitude to the same degree. Increasing degrees of breathiness and roughness contributed to progressive increases in the overall dysphonia severity. For the natural stimuli, little consistent interaction was observed between breathiness and roughness. This discrepancy in the results between synthetic and natural stimuli was speculated to result from two possible reasons. First, the process by which synthetic stimuli were created (i.e., amplitude modulation was applied after manipulating noise and the open quotient during the creation of roughness levels) would have impacted the stimuli in a manner that did not accurately capture acoustic variability in natural stimuli. This approach resulted in equal superposition of modulation on the aspiration noise as well as harmonic components of the stimuli, which may not be the case in natural voicing. Second, the execution

of the conventional 1DME tasks for each VQ on different sessions may have affected the absolute perceptual magnitudes for natural voices more than the controlled synthetic stimuli.

To accurately capture the covariance in VQ dimensions, this study introduces a novel two-dimensional magnitude estimation (2DME) task, which allows listeners to rate two VQ dimensions simultaneously and examines the differences between 1D and 2D magnitude estimates of breathiness and roughness. Although current clinical scales such as the GRBAS and CAPE-V allow for rating perceived severity of multiple VQ dimensions at the same time, they are limited by arbitrary assignment of numbers to perception and, hence, cannot accurately represent magnitude of change (e.g., Nagle, 2016, 2022; Shrivastav et al., 2005). The ordinal nature of the discrete ratings of the GRBAS scale and the continuous scale of the CAPE-V do not allow comparison of VQ judgments across time, clinicians, and patients (e.g., Stevens, 1958). For example, on a CAPE-V, a change in breathiness rating from 60 mm before voice therapy to 30 mm after voice therapy may indicate less breathiness, but the magnitude of change cannot be determined due to the inherent scale properties, just as the magnitude of change cannot be determined based on the difference between Likert ratings of 6 and 3. An accurate and precise change in VQ magnitude is vital for treatment outcome measurement of dysphonic voices, which can be obtained via ratio-level metrics using ME tasks (Eddins et al., 2021). Indeed, in prior research studies, it was shown that the use of ME ratings for hypernasality provided more consistent and reliable ratings compared to equal appearing interval scaling (Whitehill et al., 2002; Zraick & Liss, 2000).

The overall goal of this investigation was to discover and determine whether rating each of the two VQ dimensions (breathiness and roughness) independently would lead to different perceptual magnitudes than if the two were done in a combined task. This study also examined the relationship between breathiness and roughness magnitudes for each of the ME tasks. Extending the principles of selective auditory attention to VQ (Spence & Santangelo, 2010), a selective attention hypothesis would suggest that a high correlation between the 2DME and 1DME perceptual data would demonstrate that listeners can selectively attend to individual VQ dimensions in the presence of VQ covariance in a reliable manner. Although 1DME judgments are highly reliable (Eddins et al., 2021; Park et al., 2022; Shrivastav et al., 2011), even for stimuli that covary in two dimensions (Park et al., 2022), it is possible that high values on one dimension may influence ME in the other dimension. This possibility would be minimized in a 2DME task, because listeners parse the two dimensions for every judgment rather than trying to ignore one dimension and attend to the other. Thus, in a 1DME task, if the severity of the to-be-ignored dimension is high, it will exaggerate the magnitude judgment of the to-be-attended dimension. This severity exaggeration should not occur for a 2DME task when magnitudes in the two dimensions are simultaneously focused. This has been termed as severity exaggeration hypothesis in the remainder of this research note.

Method

Stimuli

To study the covariance in breathiness and roughness, a 2D matrix with varying dimensional severities was created as shown in Table 1. Specifically, for each of the focal VQ dimensions, a 4-point severity continuum was examined (none, mild, moderate, and severe) similar to the conventional GRBAS scale. The first step in sample selection involved identification of nine voices (500 ms /a/ phonations) that fit into each of the 16 severity combinations. These samples were selected through stratified-random sampling from three different disordered voice databases (Kay Elemetrics Disordered Voice Database, Sataloff/Heman Ackah [Heman-Ackah et al., 2002], & University of Florida Disordered Voice Database).[1] One additional voice was included for the moderate breathiness-normal roughness severity level. The resulting matrix with a total of 145 dysphonic voices was created by the first author (S.A.) such that samples were (a) from a wide range of laryngeal pathologies (e.g., vocal hyperfunction, presbylarynx, and paralysis), (b) from both sexes, (c) steady-state vowels with limited variation in other voice parameters (e.g., pitch, loudness, and strain),

[1]The Kay Elemetrics Disordered Voice Database was developed by the Massachusetts Eye and Ear Infirmary Voice and Speech Lab. It is commercialized by Kay Elemetrics (Kay Elemetrics Corp., 1994). It includes more than 1,400 voiced samples of sustained vowel /a/ and the first part of the Rainbow Passage. Sampling frequencies are 25 or 50 kHz. The details about Sataloff/Heman-Ackah database are provided in Heman-Ackah et al. (2002). The University of Florida Disordered Voice Database was created by recording patients from the University of Florida Ear, Nose, and Throat clinic in a quiet room (ambient or environmental noise < 40 dB measured using Type II sound level meter) using a digital audio recorder (TASCAM model) with a sampling frequency of 44100 Hz and 16-bit quantization rate. This database contains recorded samples of the vowel phonations along with read and spontaneous speech from 193 talkers with dysphonia (73 males and 120 females), resulting from various etiologies (e.g., hyperfunctional voice disorders, vocal fold paralysis, spasmodic dysphonia, and presbyphonia). Samples on the Kay Elemetrics Disordered Voice Database and Sataloff database were previously rated on a 7-point Likert scale by two trained student raters. Samples on the University of Florida Disordered Voice Database were previously rated on a 5-point Likert scale by one trained student rater and one expert rater.

Table 1. Natural dysphonic voices varying in breathiness (*x*-axis) and roughness (*y*-axis).

Roughness (R) ↑	R-Severe (RSe) B-None (BN)	R-Severe (RSe) B-Mild (BMi)	R-Severe (RSe) B-Moderate (BMo)	R-Severe (RSe) B-Severe (BSe)
	R-Moderate (RMo) B-None (BN)	R-Moderate (RMo) B-Mild (BMi)	R-Moderate (RMo) B-Moderate (BMo)	R-Moderate (RMo) B-Severe (BSe)
	R-Mild (RMi) B-None (BN)	R-Mild (RMi) B-Mild (BMi)	R-Mild (RMi) B-Moderate (BMo)	R-Mild (RMi) B-Severe (BSe)
	R-None (RN) B-None (BN)	R-None (RN) B-Mild (BMi)	R-None (RN) B-Moderate (BMo)	R-None (RN) B-Severe (BSe)
	Breathiness (B) ➔			

Note. Voices ranged across different severity levels, namely, None (N), Mild (Mi), Moderate (Mo), and Severe (Se). There was one voice in each category, resulting in a total of 16 stimuli for the perceptual testing.

and (d) devoid of pitch breaks. Three experts (authors S. A., R.S., and D.A.E.), each with over 15 years of experience in auditory-perceptual evaluation of disordered VQ, listened to the 145 stimuli during a consensus session. Stimuli were presented using Sennheiser headphones (Model HD201) in a quiet laboratory room environment. From the set of 145 voices, one stimulus was selected for each of the matrix cells in Table 1 (i.e., representing each of the VQ severity levels), resulting in a total of 16 stimuli (11 males and five females) for the perceptual experiment. Listeners in the experiments below had no metadata specific to the stimuli.

Listeners

Ten young adult listeners (two male and eight female college students) aged 20–37 years ($M \pm SD = 25 \pm 5.8$ years) were recruited from the University of South Florida to participate in this study. All participants were native speakers of American English and had hearing thresholds less than 20 dB HL via air conduction at frequencies of 250, 500, 1000, 2000, and 4000 Hz (ANSI S3.21–2004). As part of routine lab intake, the following procedures were conducted but were not used as covariates in subsequent analyses: otoscopy, tympanometry, hearing health history, noise exposure, cognitive status, and history of neurological disorders or head trauma. Participants had no or limited background in communication sciences and disorders and voice evaluation. Each participant provided informed consent in accordance with the procedures approved by the University of South Florida Institutional Review Board.

Instrumentation

Stimulus presentation and response collection were controlled by the TDT SykofizX software application (Tucker-Davis Technologies, Inc.). The stimuli were delivered monaurally via a TDT RZ6 Multi I/O processor and high-fidelity insert earphones (ER2, Etymotic Research Inc.). All samples were down-sampled to 24414 Hz to match the available sampling rate of the TDT hardware. The output level was calibrated to ensure that each stimulus was delivered at 85 dB SPL. During testing, participants were seated inside a sound-attenuating booth in front of the participant user interface that consisted of a computer monitor and a mouse for all tasks.

Procedures

After a screening session, all listeners completed the protocol shown in Table 2 to complete the perceptual experiment. Data collection for each listener was completed over approximately 6 hr and separated into three 2-hr sessions (one for 2DME and two for VQ dimension-specific 1DME tasks). All listeners completed the 2DME in their first session and the order of 1DME for breathiness and roughness were randomized across listeners.

Loudness ME Task

To familiarize listeners with the concept of ME, loudness ME training was completed prior to all experimental data collection (see Table 2). For this training, on each trial, listeners assigned a numerical value between 1 and 1,000 to indicate the loudness of a pure tone (1000 Hz, 500 ms duration). The presentation level varied in level across trials in random order from 60 dB to 92 dB SPL. Listeners were instructed to assign numbers on a ratio scale such that a sound perceived to be twice as loud as the previous sound would receive double the score. This training was chosen because loudness is a simple and clear psychophysical construct with ground truth that can be easily understood and judged in ratios or fractions by all listeners. Each stimulus was presented one time in the "practice" condition and repeated 10 times per block of trials in the "experiment" condition (see Table 2) to assess intrarater reliability. For the practice condition, experimenter visually checked the perceptual data on SykofizX software interface to verify if magnitude estimates increased from 60 to 92 dB SPL and if participants used the full 1–1,000 scale prior to the experiment condition. After providing listeners with a short break, the experiment condition data were analyzed using a custom MATLAB

Table 2. Protocol for (a) two-dimensional magnitude estimation (2DME) and (b) one-dimensional magnitude estimation (1DME) tasks.

Task order	Task	Task description and details
(a) 2DME task		
1	Loudness ME practice	9 pure tones × 1 repetition
2	Loudness ME experiment	Same 9 pure tones × 10 repetitions × 3 blocks (90 sounds per block)
3	Voice quality demo	Short PowerPoint with descriptions and audio samples of breathy, rough, and strain dimensions
4	2DME practice	16 stimuli × 1 repetition; simultaneous breathy and rough judgments
5	2DME experiment	16 stimuli × 10 repetitions × 3 blocks (160 sounds per block); simultaneous breathy and rough judgments
(b) 1DME task for breathiness or roughness		
1	Loudness ME practice	9 pure tones × 1 repetition
2	Loudness ME experiment	Same 9 pure tones × 10 repetitions × 3 blocks (90 sounds per block)
3	Voice quality demo	Short PowerPoint with descriptions and audio samples of breathy, rough, and strain dimensions
4	1DME practice	16 stimuli × 1 repetition; breathy or rough judgments
5	1DME experiment	16 stimuli × 10 repetitions × 3 blocks (160 sounds per block); breathy or rough judgments

Note. All listeners completed the 2DME in their first session and the order of 1DME for breathiness and roughness were randomized across listeners.

algorithm to ensure and verify that the (a) participants used the full range of the scale from 1 to 1,000 to depict the perceived loudness magnitude, (b) magnitude estimates increased as the dB SPL values increased (e.g., if a 64-dB SPL tone was given a numerical value of 200, tones of 72 and 92 dB SPL were given numerical values higher than 200 in an ascending manner), (c) magnitude estimates were representative of ratios or fractions, and (d) intrarater reliability across 10 stimulus repetitions was high (> .8). Following the analysis, a graphic representation with the individual repetition and average perceptual data with reliability were shown to participants to provide feedback on their performance. If participants did not meet one or more of the criteria above, additional instructions were provided, and two more blocks of the same task were completed. Most of the participants were able to complete the task with high reliability within the first two blocks.

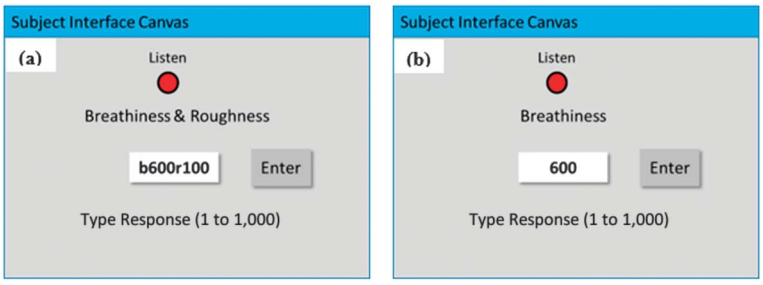
VQ Demo

Following the loudness ME task, the concept of VQ dimensions was introduced in a short slide presentation. Each of the slides introduced the individual dimensions with descriptions and audio samples that were of mild, moderate, and severe levels. The severity levels were obtained from prior ratings of the databases.[1]

2DME Task

In each trial of the 2DME task, listeners heard a voice stimulus and assigned separate values for breathiness and roughness (see Figure 1a). Here, to familiarize the listeners with the 2D rating procedure and the software interface, a "practice" condition with one repetition of the 16 stimuli was completed. Listeners were asked to use a ratio scale when assigning the magnitudes. For example, a sound perceived to be twice as breathy as another sound would be given twice its numerical magnitude. The experimenter visually checked the perceptual data on SykofizX software interface for use of the full range of scale from 1 to 1,000 as well as to see if the perceived magnitude estimates matched with the chosen severity levels of the stimuli. Following the practice, an "experiment" condition was completed with the same 16 sounds but with 10

Figure 1. Graphical user interface for (a) two-dimensional magnitude estimation (2DME) task and (b) 1DME task.



repetitions of each stimulus (16 × 10 = 160 stimuli) per block of trials. Each participant completed two more blocks of the 160 stimuli yielding 30 repetitions per stimulus. Short breaks were provided after each block. Stimuli and corresponding severity levels were randomized within and across blocks as well as across listeners. Averaging responses across 30 repetitions (16 stimuli × 10 repetitions × 3 blocks) per listener eliminated the possibility of unwanted order effects (Shrivastav et al., 2005).

1DME Task

In their subsequent sessions, listeners completed the 1DME task (see Figure 1b), evaluating breathiness and roughness VQ dimensions separately on different sessions/days. The order of the 1DME VQ tasks was counterbalanced across listeners. Similar to the 2DME task, 1DME task had "practice" and "experiment" conditions. The practice had one repetition of the 16 stimuli, and experiment had 10 repetitions of the 16 stimuli (160 stimuli) per block. A total of three blocks were completed, and short breaks were provided after each block. Responses were averaged across 30 repetitions (16 stimuli × 10 repetitions × 3 blocks) per listener. The randomization was similar to the 2DME task.

Statistical Analysis

Intraclass coefficients (ICC $_{[2,\ k]}$) were used to evaluate intra- and interlistener reliabilities. For intralistener reliability, the value of k in ICC was 30 given that each stimuli was repeated for a total of 30 times across the three blocks. For interlistener reliability, the value of k in ICC was 10, which is representative of the 10 listeners. The effects of perceived breathiness on roughness and vice versa for the 1DME and 2DME tasks were compared using iso-severity curves. The data were log transformed prior to plotting iso-severity curves to reflect the ratio scale of the ME.

To evaluate the selective attention hypothesis, four Pearson's correlations were used: two correlations to compare the relationship between the 1DME and 2DME judgments for each VQ and an additional two correlations to assess the association between perceived breathiness and roughness for each ME task. To test our severity exaggeration hypothesis regarding possible differences between the results of 1DME and 2DME under the presence of high levels of covarying VQs, paired *t*-test analyses, as shown in Table 3, were completed. Stimuli from normal and mild severity levels were categorized into a "low" stimulus subset, while stimuli from the moderate and severe levels were categorized into a "high" stimulus subset (see Table 3). To determine whether the influence of covarying roughness levels on perceived breathiness is different between the ME tasks, paired *t* tests were conducted on perceived breathiness magnitudes, separately for low and high roughness stimulus subsets. Similarly, to determine whether the influence of covarying breathiness levels on perceived roughness is different between the ME tasks, paired *t* tests were conducted on perceived roughness magnitudes, separately for the low and high breathiness stimulus subsets. In total, four paired *t* tests were conducted to evaluate the severity exaggeration hypothesis. The significance level was adjusted to .006 with Bonferroni correction.

Results

For the 1DME and 2DME tasks, intra- and interlistener reliability was high when estimating the magnitude of breathiness and roughness, as shown in Table 4. The effects of the roughness severity level on perceived breathiness and the effects of breathiness severity level on perceived roughness were analyzed using iso-severity curves as shown in Figures 2 and 3. Iso-severity curves indicated that only one breathiness level (B–N) had prominent positive slope as shown in Figure 2b (1DME). The B–N samples with higher roughness were judged to have greater perceived breathiness relative to the B–N samples with lower roughness. For the mild, moderate, and severe breathiness samples (B-Mi, Mo, and Se), variations in roughness did not alter breathiness magnitude estimates

Table 3. Paired *t*-test analysis comparisons to evaluate severity exaggeration hypothesis.

Stimulus set	Stimuli	Independent variable	Dependent variable	*t*	*p*
Low roughness (eight stimuli)	RN + BN, BMi, BMo, BSe RMi + BN, BMi, BMo, BSe	1DME vs. 2DME	Perceived breathiness magnitudes	1.47	.15
High roughness (eight stimuli)	RMo + BN, BMi, BMo, BSe RSe + BN, BMi, BMo, BSe	1DME vs. 2DME	Perceived breathiness magnitudes	5.41	< .001**
Low breathiness (eight stimuli)	BN + RN, RMi, RMo, RSe BMi + RN, RMi, RMo, RSe	1DME vs. 2DME	Perceived roughness magnitudes	−1.52	.13
High breathiness (eight stimuli)	BMo + RN, RMi, RMo, RSe BSe + RN, RMi, RMo, RSe	1DME vs. 2DME	Perceived roughness magnitudes	3.11	.003**

Note. Asterisks (**) indicate significance. R = roughness; N = None; B = breathiness; Mi = Mild; Mo = Moderate; Se = Severe; 1DME = one-dimensional magnitude estimation; 2DME = two-dimensional magnitude estimation.

Table 4. Intra- and interlistener reliability computed via ICC (2, k, absolute agreement).

Reliability type/task	Intralistener reliability	Interlistener reliability
	M (range)	*M*
Ratings of breathiness on 2DME	.97 (.94–.99)	.97
Ratings of roughness on 2DME	.96 (.92–.99)	.94
Ratings of breathiness on 1DME	.99 (.97–.99)	.98
Ratings of roughness on 1DME	.97 (.92–.99)	.94

Note. ICC = intraclass coefficient; 2DME = two-dimensional magnitude estimation; 1DME = one-dimensional magnitude estimation.

uniformly. In contrast, as shown in the roughness contours of Figures 3a (2DME) and Figures 3b (1DME), the perceived roughness magnitudes remain relatively similar within the same roughness group despite changes in breathiness.

To evaluate the selective attention hypothesis, Pearson's correlations were computed. As shown in Figures 4a and 4b, there was a strong correlation between 2DME and 1DME judgments for each VQ dimension ($p < .001$), indicating that listeners are highly reliable in judging individual VQ dimensions in the presence of other VQ dimensions. Additionally, the correlation between perceived breathiness and roughness in the 1DME task was low and nonsignificant ($r = .40$, $p = .14$). There was also no significant correlation between perceived breathiness and roughness magnitudes in the 2DME task ($r = .01$, $p = .97$).

Figure 2. Log-transformed breathiness magnitudes ($M \pm 95\%$ CI) as a function of roughness level (RN to RSe) for each breathiness level (BN to BSe). CI = confidence interval; B = breathiness; R = roughness; N = none; Mi = mild; Mo = moderate; Se = severe. (a) Ratings from two-dimensional magnitude estimation (2DME) and (b) ratings from one-dimensional magnitude estimation (1DME). Asterisks indicate significant positive slope ($p < .05$).

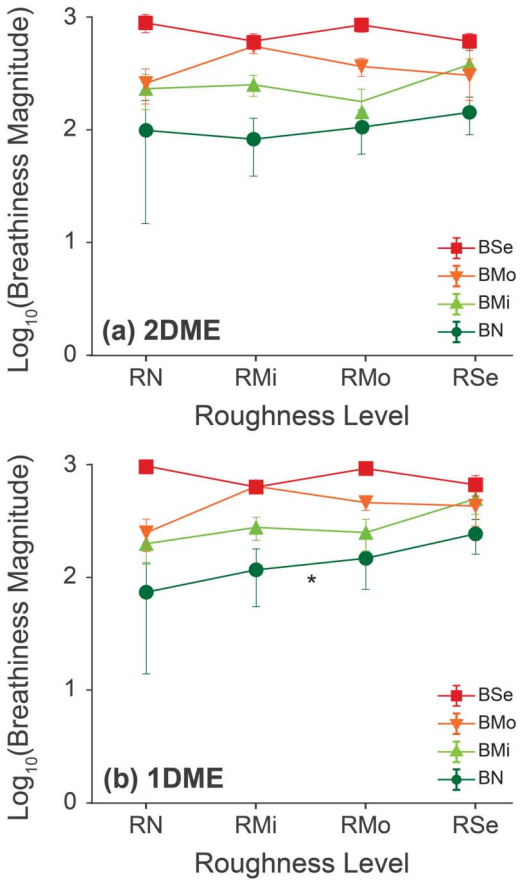


Figure 3. Log-transformed roughness magnitudes ($M \pm 95\%$ CI) as a function of breathiness level (BN to BSe) for each roughness level (RN to RSe). B = breathiness; R = roughness; N = none; Mi = mild; Mo = moderate; Se = severe. (a) Ratings from two-dimensional magnitude estimation (2DME) and (b) ratings from one-dimensional magnitude estimation (1DME).

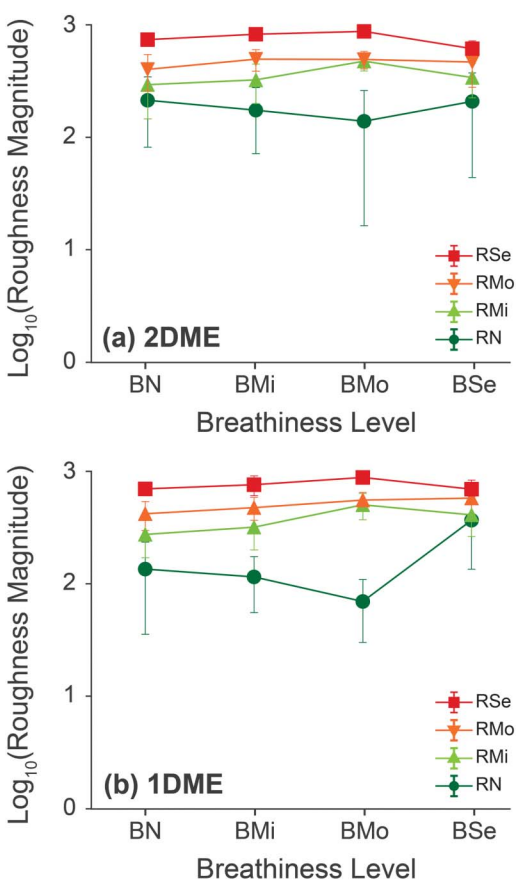
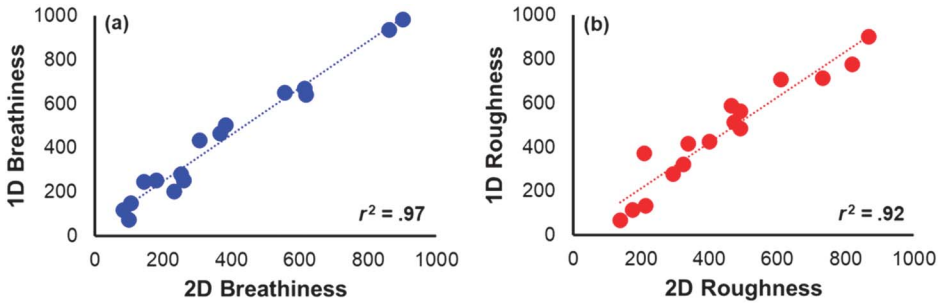


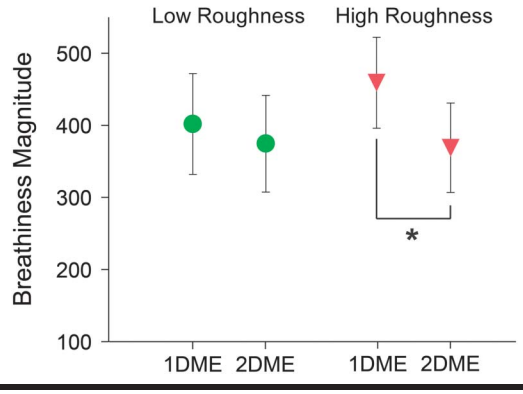
Figure 4. Relationship between perceived breathiness (a) and roughness (b) on a two-dimensional magnitude estimation (2DME) versus one-dimensional magnitude estimation (1DME) task.



These results support the selective attention hypothesis demonstrating the independence of VQ dimensions when evaluated using the robust 2DME task.

To evaluate the severity exaggeration hypothesis, paired *t* tests were conducted. The results showed that, for stimuli with a high level of roughness, perceived breathiness magnitudes were significantly greater ($t = 5.41$, $p < .001$) in the 1DME than in the 2DME task, as shown in Figure 5. However, for stimuli with a low level of roughness, there was no significant difference in the magnitude of breathiness between the two tasks ($p = .15$). This statistical outcome confirms our severity exaggeration hypothesis and demonstrates the superiority of the 2DME task over the 1DME tasks in evaluating the magnitude of VQ in dysphonic samples that covary in more than one VQ dimension. Similarly, for stimuli with a high level of breathiness, the perceived roughness magnitude was significantly greater ($t = 3.11$, $p = .003$) in the 1DME task than in the 2DME task (see Figure 6). However, for stimuli with a low level of breathiness, there was no significant difference in roughness magnitude between the two tasks ($p = .13$).

Figure 5. Breathiness magnitudes ($M \pm 95\%$ CI) for one-dimensional magnitude estimation (1DME) and two-dimensional magnitude estimation (2DME) tasks in samples with low (green) and high (red) roughness. The asterisk indicates significant difference ($p < .001$).

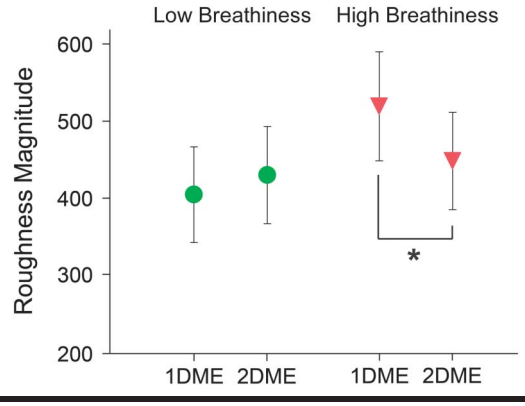


Discussion

Only a limited number of studies have examined the interaction between multiple VQ dimensions. In this exploratory study, we explicitly evaluated how the covariance of breathy and rough voice qualities in natural dysphonic voices may affect VQ judgments. To do so, the VQ judgments from a novel 2DME task were compared to those obtained using a conventional 1DME task. Careful selection of natural dysphonic voices that varied along the severity continuum for two specific VQ dimensions (breathiness and roughness) was essential to sample the range of dysphonia severity observed in patients.

The strong listener reliability and high correlations between the 2DME and 1DME estimates indicate the robust nature of the novel 2DME task. In addition, although the stimuli covaried in both VQ dimensions, the low correlation values between breathiness and roughness from both tasks support breathiness and roughness as being two distinct VQ dimensions that listeners can attend

Figure 6. Roughness magnitudes ($M \pm 95\%$ CI) for one-dimensional magnitude estimation (1DME) and two-dimensional magnitude estimation (2DME) tasks in samples with low (green) and high (red) breathiness. The asterisk indicates significant difference ($p = .003$).



to selectively. Consistent with the selective attention hypothesis, listeners were able to distinguish breathiness from roughness when performing both the 1DME and 2DME tasks, as evidenced by the strong correlations between the quality-specific values obtained from both methods.

General trends from the iso-severity curves show that the presence of roughness in a voice had a limited influence on breathiness judgments, and judgments of roughness did not seem to be impacted by low levels of breathiness in a dysphonic voice. However, for voices with higher levels of breathiness or roughness, perceptual judgments on one dimension can be significantly influenced by another, especially in the 1DME task. The perceived magnitude of each of the two VQ dimensions studied here was higher in the 1DME task than in the 2DME task, but only for stimuli with the highest levels of severity along one of the VQ dimensions. This suggests that the assigned VQ magnitude in a 1DME task is biased by the covariation of VQ along the second dimension. A comparison of VQ judgments obtained through the 1DME and 2DME tasks shows that this effect of covarying VQ can be minimized in 2DME, likely due to listeners' simultaneous focus on both quality dimensions. Therefore, these results provide empirical support for the severity exaggeration hypothesis, along with data suggesting that the 2DME task is more resilient to exaggeration of the severity of one dimension by the severity of another.

As with any experiment of this nature, the results were constrained by the nature of the stimuli tested and the limits of the psychophysical task(s) used to elicit judgments of perceptual magnitude. Here, the stimulus continua were identified through a stratified sampling approach to identify a small set of stimuli that captured the wide range of breathiness and roughness observed in dysphonic voices. However, natural stimuli vary along multiple dimensions, such as pitch (correlated with fundamental frequency) or vowel characteristics (correlated with formant frequencies), which may have added some variance to the resulting data. Another consideration is the potential bias inherent to the ME task (Patel et al., 2010). For example, in a 1DME task, perceptual judgments are significantly affected by the number of stimuli and the overall range of the percept under study. Similar biases may also occur in the 2DME task and may impact generalization, particularly cross-study comparisons where the range of roughness and breathiness across different experiments may not be directly comparable. All the listeners completed each of the ME sessions (one 2DME and two dimension-specific 1DME experiments) across different days for the "order effect" to be minimized. The order of VQ dimension–specific 1DMEs were randomized across listeners. All the stimuli were also randomized within and across the three blocks, across the ME sessions per listener, and across listeners. Despite these controls, the completion of 2DME first and then 1DME may be perceived as a limitation. Finally, this study primarily focused on breathiness and roughness and did not consider the third primary VQ dimension of strain. A future study examining dysphonic voices that vary along the continuums of breathy–strain, rough–strain, and breathy–rough–strain is warranted. Future studies on connected speech are also warranted.

Although several investigations have intentionally constrained test stimuli for experimental purposes, the development of robust clinical tools for VQ evaluation requires us to measure, understand, and model the variability seen in natural voices. Over the past several years, significant advances have been made in quantifying changes in dysphonic VQ in a precise and consistent manner (Anand, 2023; Anand et al., 2019; Eddins et al., 2021; Park et al., 2022, 2023; Patel et al., 2012a, 2012b; Shrivastav et al., 2005). The primary goal of using the ME task in the current investigation was to tease out possible covariance among voice qualities at a granular level, and the outcomes of this experiment indeed help to better understand the underlying covariance in breathiness and roughness—a frequent occurrence in dysphonic voices—and support the practice of evaluating multiple dimensions simultaneously using a 2D or 3D perceptual task. However, the ME task itself likely has limited utility as a clinical assessment procedure. In fact, the authors are unaware of any clinical protocols that use the ME task. In addition to our current use of the ME task to explore covariance, the ME task also serves as a critical procedure in developing standard perceptual scales of VQ suitable for clinical use (Eddins et al., 2021).

The development of standard perceptual scales for use in quantifying VQ (i.e., scale of perceived breathiness, roughness, or strain) can be based on the methods used to develop other perceptual scales such as the Sone scale of loudness (Stevens, 1936). It is essential to establish the relationship between physical units associated with the VQ (using a matching task) and the perceived magnitude (using a ME task). This mapping of perceptual quantities to physical quantities is the basis for a perceptual scale. To facilitate the usability of such scales, a physical unit in the middle of the continuum of relevant perceptual values is chosen as the standard reference point on the scale. All other points on the perceptual continuum are rescaled relative to this standard reference value. The perceived VQ is then judged relative to this one scale unit. Once developed, the scale supports simple, easy to use perceptual judgments as well as computational methods to estimate perceptual magnitude. The development of standard scales renders perceptual VQ measurements operable in clinical

practice contexts unlike the cumbersome matching and ME tasks. Importantly, the resulting scales support extremely desirable measurement properties such as the ability to quantify magnitude of change.

Data Availability Statement

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

Acknowledgments

This work was supported by the National Institute on Deafness and Other Communication Disorders R01DC009029 (D.A.E. and R.S.).

References

American National Standards Institute. (2004). *Methods for manual pure-tone threshold audiometry* (ANSI S3.21–2004).

Anand, S. (2023). Perceptual and computational estimates of vocal breathiness and roughness in sustained phonation and connected speech. *Journal of Voice*. Advance online publication. https://doi.org/10.1016/j.jvoice.2023.02.014

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

American Speech-Language-Hearing Association. (2002). *Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V)*.

Baldner, E. F., 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

Dejonckere, P. H. (1995). Principal components in voice pathology. *Voice, 4,* 96–105.

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., Anand, S., Lang, A., & Shrivastav, R. (2021). Developing clinically relevant scales of breathy and rough voice quality. *Journal of Voice, 35*(4), 663.e9–663.e16. https://doi.org/10.1016/j.jvoice.2019.12.021

Gescheider, G. (1976). *Psychophysics: Method and theory*. Erlbaum.

Green, D. M., Luce, R. D., & Duncan, J. E. (1977). Variability and sequential effects in magnitude production and estimation of auditory intensity. *Perception & Psychophysics, 22*(5), 450–456. https://doi.org/10.3758/BF03199510

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., 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

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

Kay Elemetrics Corp. (1994). Disordered voice database model 4337.

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)

McKenna, F. P. (1985). Another look at the 'new psychophysics.' *British Journal of Psychology, 76*(1), 97–109. https://doi.org/10.1111/j.2044-8295.1985.tb01934.x

Nagle, K. F. (2016). Emerging scientist: Challenges to CAPE-V as a standard. *Perspectives of the ASHA Special Interest Groups, 1*(3), 47–53. https://doi.org/10.1044/persp1.SIG3.47

Nagle, K. F. (2022). Clinical use of the CAPE-V scales: Agreement, reliability and notes on voice quality, reliability and notes on voice quality. *Journal of Voice*. Advance online publication. https://doi.org/10.1016/j.jvoice.2022.11.014

Park, Y., Anand, S., Gifford, S. M., Shrivastav, R., & Eddins, D. A. (2023). Development and validation of a single-variable comparison stimulus for matching strained voice quality using a psychoacoustic framework. *Journal of Speech, Language, and Hearing Research, 66*(1), 16–29. https://doi.org/10.1044/2022_JSLHR-22-00280

Park, Y., Anand, S., Kopf, L. M., Shrivastav, R., & Eddins, D. A. (2022). Interactions between breathy and rough voice qualities and their contributions to overall dysphonia severity. *Journal of Speech, Language, and Hearing Research, 65*(11), 4071–4084. https://doi.org/10.1044/2022_JSLHR-22-00012

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)

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., & 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)

Spence, C., & Santangelo, V. (2010). Auditory attention. In C. Plack (Ed.), *Oxford handbook of auditory science: Hearing* (pp. 249–270). Oxford University Press.

Stevens, S. S. (1936). A scale for the measurement of a psychological magnitude: Loudness. *Psychological Review, 43*(5), 405–416. https://doi.org/10.1037/h0058773

Stevens, S. S. (1956). The direct estimation of sensory magnitudes: Loudness. *The American Journal of Psychology, 69*(1), 1–25. https://doi.org/10.2307/1418112

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review, 64*(3), 153–181. https://doi.org/10.1037/h0046162

Stevens, S. S. (1958). Problems and methods of psychophysics. *Psychological Bulletin, 55*(4), 177–196. https://doi.org/10.1037/h0044251

Stevens, S. S. (1959). Measurement, psychophysics, and utility. In C. W. Churchman & P. Ratoosh (Eds.), *Measurement: Definitions and theories*. Wiley.

Walden, P. R., & Rau, S. (2022). Individual voice dimensions' prediction of overall dysphonia severity on two auditory-perceptual scales. *Journal of Speech, Language, and Hearing Research, 65*(8), 2759–2777. https://doi.org/10.1044/2022_JSLHR-21-00689

Whitehill, T. L., Lee, A. & Chun, J. C. (2002). Direct magnitude estimation and interval scaling of hypernasality. *Journal of Speech, Language, and Hearing Research, 45*(1), 80–88. https://doi.org/10.1044/1092-4388(2002/006)

Zraick, R. I., & Liss, J. M. (2000). A comparison of equal-appearing interval scaling and direct magnitude estimation of nasal voice quality. *Journal of Speech, Language, and Hearing Research, 43*(4), 979–988. https://doi.org/10.1044/jslhr.4304.979