Developing Clinically Relevant Scales of Breathy and Rough Voice Quality

***David A. Eddins, *Supraja Anand, †Arianna Lang, and ‡Rahul Shrivastav,** **Tampa, Florida, †Kyle, Texas, and ‡Athens, Georgia*

Summary: The most common measurement tools used in the perceptual evaluation of voice quality yield ordinal data and thus do not support the establishment of mathematical relationships among different measurement values. This makes their interpretation challenging. Among the many desirable features of any psychophysical measurement tool is the ability to quantify the difference between two or more measurements and the ability to interpret the measurements in a manner that is related to the experience of the observer. The former allows one to compare among measurements using simple mathematics, while the latter allows that comparison to be interpreted in constructive ways. In this paper we describe the development of standard measurement scales for two dimensions of voice quality, following an approach that has been applied successfully to the perception of loudness. The scales follow step-by-step procedures used to develop the sone scale of loudness, which ties physical units to the perceptual estimates of loudness magnitude. Goals of the current work include development of analogous scales for the perception of breathy and rough voice qualities. First, the relationship between perceived voice quality and physical units were established using single-variable matching tasks. Second, the relationship between a change in physical units from the single-variable matching tasks and perceived voice quality magnitude were established using magnitude estimation tasks. Third, single reference points were identified on breathy and rough continuums. Finally, all points on the newly established voice quality continuums were rescaled relative to these arbitrary reference points. The proposed breathiness and roughness scales result in ratio-level data with standard measurement units that support quantitative comparisons of perceptual judgments. Such judgments can be used, for example, to compare magnitude of change pre- and post-treatment.

Key Words: Voice quality−Clinical scales−Breathy voice−Rough voice.

INTRODUCTION

Quantifying voice quality or changes in voice quality in the context of a clinical evaluation of voice can be challenging both for practical and theoretical reasons. Nevertheless, recovery of a voice over time and/or with treatment typically is associated with improved quality-of-life and it is expected that, successful treatment practices result in positive changes in voice quality. Thus, there is a need for accurate quantification of improvement or decline in voice quality. Current clinical practice nearly always relies on subjective assessment, using simple but limited methods, and frequently includes quantitative measures of the vocal acoustic signal, using algorithms available in commercial hardware/software systems. While easy to use, tools used to measure voice quality in clinical settings lack the psychometric qualities of perceptual tasks used in other clinical settings such as audiology and neuropsychology. A long-term goal of the work reported here is to improve the clinical and research tools available for assessing dysphonic voice quality. One way to improve the measurement of voice quality in clinical and laboratory settings is to develop standard measurement scales for fundamental voice qualities such as breathiness and roughness. That is the focus of the current study.

The most common clinical tools used for evaluating voice quality consist of auditory-perceptual assessments such as the Grade, Rough, Breathy, Asthenia, Strain scale[1] and the Consensus Auditory-Perceptual Evaluation of Voice.[2] Although these are considered the current "gold standards," there are two fundamental weaknesses in the rating-scale approach: the ordinal nature of the task and poor rater reliability as observed empirically. The ordinal nature of rating-scale measures eliminates the possibility of comparing voice quality measurement values (eg, by addition, subtraction, or division) across time, clinicians, and patients. For example, using visual-analog scale of the Consensus Auditory-Perceptual Evaluation of Voice, a change in roughness rating from 80 mm to 40 mm after vocal rest may represent less roughness, but the magnitude of the change cannot be determined because of inherent ordinal scale properties. Ratio-level data permit such comparisons. Partial sacrifice of psychometric properties with clinically useful tools is expected, but the ability to compare magnitude via ratio-level metrics is essential to outcome measurement.

With rating-scale measures, listener reliability can be inadequate due to various factors including the type of rating scale used, nuances of the voice being evaluated, stimulus context, and environmental factors.[3−5] In some investigations, reliability is reported to be quite high[6−7] while others report considerably lower reliability.[8−10] Reliability often varies by quality dimension as well, with rater reliability for breathiness generally being a bit higher than for roughness and both of those

Accepted for publication December 23, 2019.
Funding: Work supported by NIH NIDCD R01 DC009029.
From the *Department of Communication Sciences and Disorders, University of South Florida, Tampa, Florida; †120 Peach Tree Pass, Kyle, Texas; and the ‡Office of the Vice President for Instruction, University of Georgia, Athens, Georgia.
Address correspondence and reprint requests to David A. Eddins, Department of Communication Sciences and Disorders, University of South Florida, 4202 E. Fowler Avenue, PCD 1017, Tampa, FL 33620. E-mail: deddins@usf.edu
Journal of Voice, Vol. 35, No. 4, pp. 663.e9−663.e16
0892-1997

https://doi.org/10.1016/j.jvoice.2019.12.021

considerably higher than for ratings of strain.[11] When repeated ratings of the same voice are obtained for the same judge, later ratings may be different than the initial ratings. Ratings may differ between clinical and laboratory settings, where clinical bias may exacerbate ratings in the clinic.[12] While some experts may consider the reliability of clinical perceptual voice evaluations based on ratings to be acceptable, and others may consider methods to increase reliability,[13–16] accepting poor reliability or the use of procedures to improve reliability does not overcome the fundamental limitations of an ordinal scale.

The problems associated with measuring voice quality are not unique to voice and are not new. Such problems are encountered when trying to measure any qualitative sensory attribute. Methods developed to quantify the loudness of sound, for example, can serve as a model for how one might deal with accurate estimates of the perception of breathy or rough voice quality. Stevens (1936)[17] noted perceptual scales are needed "to facilitate the description of natural phenomenon in terms of functional relationships using the symbols of conventional mathematics." He pointed out that when quantifying perception, it is insufficient to use ordinal scales such as A to F or 1 to 7. Rather, Stevens argued that it is desirable to use numbers that have an obvious, simple relationship to the perception of the observer. An important attribute of such a scale is that it should be able to connote the relative magnitude of the stimuli to be represented by the scale. The sone scale developed by Stevens (1936)[17] is an example of a magnitude-based psychophysical scale that correlates sound intensity levels in dB to the sensation of loudness. The sone is a unit of perceived loudness. Doubling the perceived loudness doubles the sone value. Stevens (1936)[17] suggested that a scale could be constructed if "a number *N* was assigned to a particular magnitude, and the number *N/2* would be assigned to the magnitude which appears half as great to the experiencing individual." Therefore, the scale was designed, through various experiments, so that 1 sone unit was defined as the perceived loudness of a 1000 Hz tone with an intensity of 40 dB above absolute threshold. A 47 dB tone was judged to be 2 sones because it produced a sound perceived to be twice as loud as the 1 sone reference. The sone scale, developed over 80 years ago, is still used widely to quantify the loudness of arbitrary simple and complex sounds such as the loudness of room fans, refrigerators, or automobile cabin noise. Analogous scales can be developed to quantify vocal breathiness and roughness.

Following the methods of Stevens used in the context of the development of the sone scale, a similar process will be undertaken here to produce scales for the dysphonic voice qualities of breathiness and roughness. The process involves four basic steps outlined by Stevens (1936).[17] Step one involves the establishment of *physical scales* that express the relationship of the magnitude of the perceived voice quality to appropriate physical units. In the case of loudness, a loudness matching task was used. Step two involves establishment of *psychophysical scales* that express the relationship between a change in perceived magnitude to a change in physical units, effectively determining the relationship among different perceptual magnitudes. Stevens used a loudness magnitude estimation (ME) task for this step. Step three defines a single reference point on the psychophysical scale. In the case of loudness, a 40-dB SPL, 1000 Hz tone was established as the reference and defined as having a loudness of 1 sone. Step four involves a rescaling of all other points on the psychophysical scale continuum relative to the newly defined reference point (eg, a loudness less than or greater than 1 sone). Following these four steps, the goal, through the use of analogous scales for vocal breathiness and vocal roughness, is to increase the utility of voice quality assessment in the clinic and thus provide a precise, ratio-level, standardized measurement for voice quality dimensions.

For step one in the development of scales for breathy and rough voice quality, we lean on previous research using a single-variable matching task (SVMT) in which listeners matched the perceived dysphonic voice quality to the perceived quality of a synthetic comparison (nonspeech) sound that had a single adjustable parameter that increased or decreased the perceptual quality to be matched. In the case of vocal breathiness, the variable parameter of the comparison stimulus was the signal-to-noise ratio (SNR) with units of dB. Adjustment of this parameter makes the comparison sound more or less "noisy," as shown in the left panel of Figure 1. For the roughness quality, the parameter of the comparison manipulated was amplitude modulation depth in dB, as shown in the right panel of Figure 1.

Adjustment of this parameter makes the comparison sound more or less "rough." Thus, the previous matching experiments[5,18] provide physical units that correspond to perceived voice qualities of breathiness (Figure 2, left panel) and roughness (Figure 2, right panel).

In a clinical setting, one does not expect a clinician to perceive breathiness in terms of the absolute SNR in dB or to perceive a change following treatment or pathology progression as a change in dB SNR. Such numerical estimations would be difficult, at best, even for a trained professional. Furthermore, the matching task used to generate the data in Figure 2 is not practical for use in a clinical setting. The method does, however, map perception to physical units. Step two in developing perceptual scales is to determine the relationship between the change in physical units in each voice quality dimension (SNR or amplitude modulation depth in dB) and the change in perceived magnitude. The experiment reported in this investigation accomplishes step two. In this experiment, breathy and rough voice quality attributes will be gauged through a ME task. Quality judgments will be mapped to the matching stimuli of step one using a wide range of physical values representing breathiness in dB SNR. Similar procedures will be taken to map perception to physical units representing roughness in dB amplitude modulation (AM) depth. In step three, a single reference point will be defined on the VQ continuum (eg, analogous to establishing 40 Phons as a standard reference point for loudness). In step four, all perceptual data will be rescaled relative to that single reference and the reference value will be considered one scale unit. The quality of any voice will be compared to that reference value of one-scale unit and

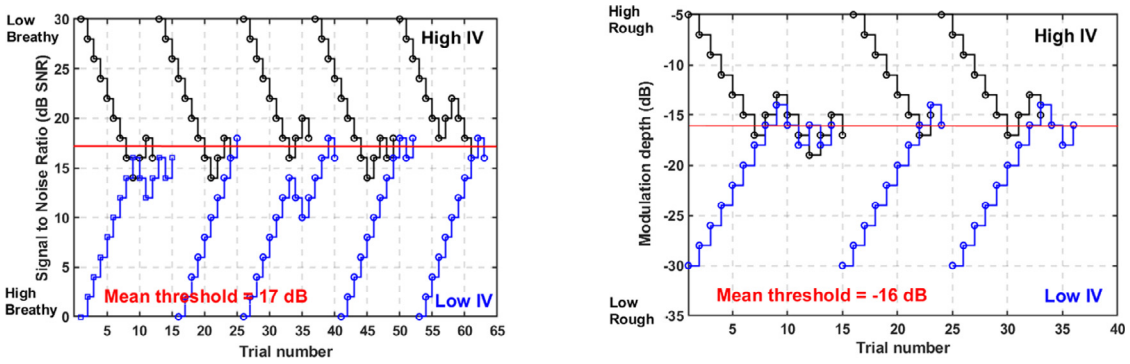

FIGURE 1. Left: Breathiness matching procedure with SNR in dB (y axis) across trials (x axis). Listener's response is depicted for a single talker/dysphonic voice using the up-down tracking procedure in SVMT. The first trial of each adaptive track started with an initial value (IV) of either 30 dB (low breathy) or 0 dB (high breathy), and results for the high- and low-IV are averaged for each of five replicates of each stimulus. Right: Roughness matching procedure with modulation depth in dB (y axis) across trials (x axis). Listener's response is depicted for a single talker/dysphonic voice using the up-down tracking procedure in SVMT. The first trial of each adaptive track started with an IV of either −35 dB (low rough) or −5 dB (high rough), and results for the high- and low-IV are averaged for each of five replicates of each stimulus.

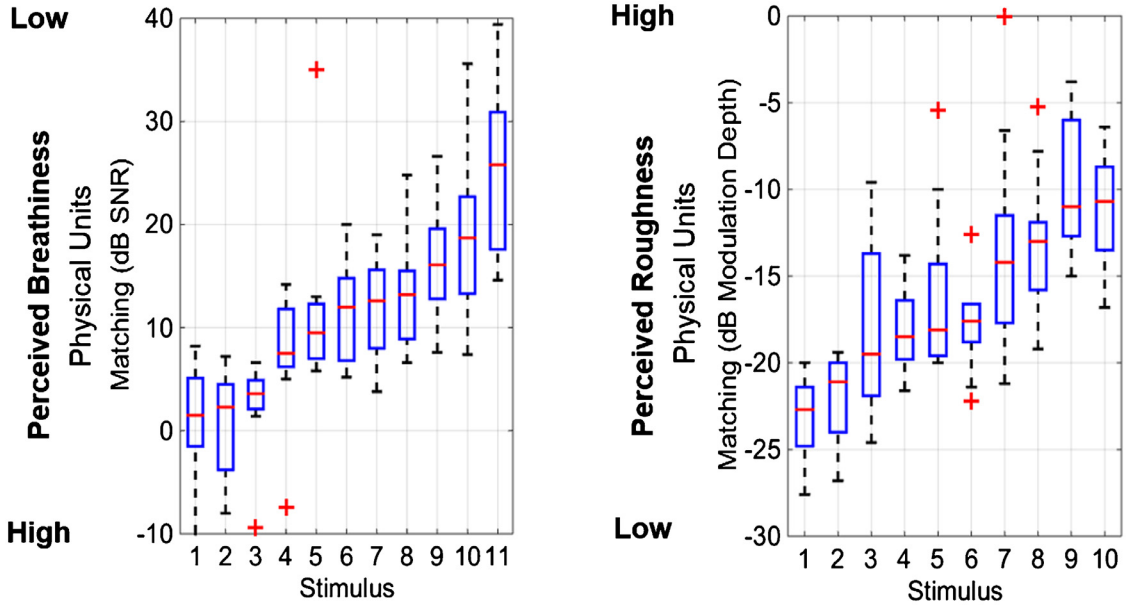

FIGURE 2. Left: Breathiness matching values for 11 natural voice stimuli ranging from normal to severely breathy. Matching values are expressed in physical units of dB signal-to-noise ratio, SNR (Adapted from Patel et al, 2010). Right: Roughness matching values for 10 natural voice stimuli ranging from normal to severely rough (adapted from Shrivastav and Eddins, 2012). Matching values are expressed in physical units of dB amplitude modulation depth. In each panel, the box and whiskers plots represent data from 10 listeners. See Methods for more details regarding comparison stimuli.

expressed numerically as a fraction (eg, $^1/_2$ or 0.5 scale units) or multiple (e.g., 2 times or 2 scale units) relative to the reference value. In this manner, the new scale is considered to be a ratio scale, in that the quality of the voice sample is judged as a ratio of the reference value. This also is analogous to the common decibel scale where the reference is in units of pressure or intensity rather than modulation depth or SNR. The mathematical advantages of a ratio-level scale has been championed by many authors.[19–22]

Such quantification would support quantification of change over the course of pathology progression, before, during, and after treatment, and across treatment methods. Consequently, there is a need to develop a simple and relatively quick measurement tool for voice quality perception that can be effectively translated to the clinic but has strong psychometric properties. Here we report the development of such scales for the breathy and rough voice qualities.

METHOD

Listeners

Twenty-five listeners ranging in age from 19 to 51 years were recruited from the University of South Florida to participate in this study. All listeners had normal hearing bilaterally

(pure tone thresholds <25 dB HL 250 to 8000 Hz; ANSI, 2010)[23], confirmed through a hearing evaluation, and were native speakers of American English. Participants had minimal to no previous exposure to dysphonic voice quality. All listeners consented to participate according to procedures approved by the university institutional review board.

Stimuli

Practice Stimuli: Nine pure tone stimuli were created for the loudness training task. Each stimulus consisted of a 1000 Hz tone, 500 milliseconds in duration including 20 milliseconds cosine-squared rise/fall ramps, and ranging in level from 60 to 92 dB SPL in nine 4-dB steps.

Experimental Stimuli: The stimuli used in the breathiness arm of this study are identical to the comparison stimuli used the matching task of previous experiment to gauge breathiness.[24] Those stimuli were based on the acoustic characteristics of a set of disordered voice stimuli from the Kay Elemetrics Disordered Voice Database (KEDVD; Kay Elemetrics, Inc., Lincoln Park, New Jersey). Accordingly, the breathy comparisons were created by mixing a sawtooth wave and noise to create a carrier stimulus. The spectrum and the harmonics of the sawtooth wave are associated with the glottal source of natural speech, therefore providing a close approximation to the perception of periodic voiced speech. The sawtooth wave had a fundamental frequency (f_0) of 151 Hz mixed with speech-shaped noise. The sawtooth and noise were low-pass filtered to have a slope of −7 dB per octave above 151 Hz.

Thirteen breathy stimuli were created by varying the SNR of the carrier over a range of 0 to 36 dB in 3 dB steps. This range was chosen to exceed the range established in previous matching experiments.[24] A stimulus with 0 dB SNR is said to have extremely high breathiness, while one with 36 dB SNR would indicate extremely low breathiness. These stimuli exceed the range of breathiness observed in natural stimuli that span the range from normal to severely dysphonic voices and the relationship between those voices and these synthetic stimuli was established in the earlier studies.

The stimuli for the roughness arm of the study were created similar to prior matching experiment on roughness.[25] The sawtooth plus noise stimulus used in breathiness experiments was fixed in SNR (+12 dB). This carrier stimulus was used to create 13 stimuli that varied in degree of roughness by imposing amplitude modulation on the carrier. As other experiments have studied, the perceptual roughness of a sound can be produced by amplitude modulating sounds.[26,27] The degree of modulation (modulation depth) directly impacts the degree of perceived roughness.[28,29] The modulating waveform took the form:

$$Y(t) = 1 + m[\sin(2\pi f t_+ T)]^4 * c(t)$$

where m is the modulation depth (0 to 1), f is modulation frequency in Hz (fixed at 25 Hz), t is time in seconds, T is starting phase (fixed at 0 radians), and c is the sawtooth-plus-noise carrier. The modulation depth is typically varied on a logarithmic scale (in dB) where the modulation index equals $20*\log_{10}(m)$, and m can vary from 0 to 1.The stimuli for this experiment varied in AM depth from −12 to −36 dB in 2 dB steps spanning the range of relevant depths matched to stimuli from extreme to limited perceived roughness.[28] A stimulus with an AM depth of −12 dB is said to have extremely high roughness, while one with an AM depth of −36 dB would indicate extremely low roughness.

Instrumentation

All experimental procedures were conducted in a sound attenuating chamber. Stimulus generation, presentation, and response collection were controlled via Tucker-Davis Technologies (TDT) SykofizX software via TDT System 3 RZ6 real-time processor connected to a TDT HB6 headphone buffer and stimuli were delivered to the listener monaurally at a level of 80 dB SPL through Etymotic Research Inc. ER2 insert earphones. The listener interface consisted of a flat-screen display, keyboard, and mouse. The interface displayed a red visual cue to prompt that the stimulus was being presented. Participants then typed a numeric perceptual rating in a white box displayed on the screen followed by the "Enter" key once they had rated the stimulus.

Procedures

The breathy and rough voice quality attributes were evaluated through a direct ME task without anchors. For breathy voice quality, breathiness judgments were mapped to physical units representing breathiness in terms of SNR in dB. For rough voice quality, roughness judgments were mapped to physical units representing roughness in terms of AM depth in dB. The ranges of breathiness units and roughness units were based on previous published studies of those dimensions using a SVMT.

Magnitude estimation for loudness (practice)

Prior to data collection for the breathiness and roughness conditions, participants were familiarized with the task by completing a ME practice task for the loudness of pure tone stimuli presented at 9 different intensity levels. The participants were instructed to estimate the loudness of the tone at each level on a ratio scale of 1 to 1000 and to enter that value on the user interface by typing using a wireless keyboard. A value of 1 indicated extremely low loudness while a value of 1000 indicated extremely high loudness. Data were collected and analyzed for this task before proceeding to the breathiness and roughness ME tasks to assure that the participant understood the use of the interface, to encourage the participant to use a wide range of magnitudes in their judgments on the ratio-level ME scale, and to evaluate whether or not the participant could accurately and consistently order pure tones according to loudness.

Magnitude estimation for breathy and rough stimuli (experiment)

A similar ME task examined the perceived roughness and breathiness of each element in a set of 13 stimuli described

above as spanning the relevant range of perceived vocal breathiness or roughness over natural and dysphonic voices. For the breathiness and roughness conditions, each of the 13 stimuli was presented 10 times in random order. The test block for each of the breathiness and roughness conditions consisted of three separate runs (13 stimuli × 10 repetitions × 3 runs). The participants were instructed to estimate the magnitude of breathiness/roughness on a ratio scale of 1 to 1000 where a value of 1 indicated extremely low breathiness/roughness and a value of 1000 indicated a sound with extremely high breathiness/roughness. The participant was told that the roughness quality could also be judged by the amount "fluctuation strength" and that breathiness could be judged as the overall "noisiness" of the sound. To assist the participant in understanding the use of a ratio scale, it was explained that a sound perceived to be twice as rough/breathy as the previous sound should be given double the score. A sound perceived to be 10 times as breathy/rough as the previous sound should be given a score 10 times as great as the previous sound. Likewise, a sound that is perceived to be ¼ as breathy/rough as the previous sound should be given a score that is ¼ the value of the previous sound.

For each of the breathiness and roughness conditions, a single practice block preceded the main task. In the practice blocks, three (rather than 10) repetitions of each stimulus were presented for three separate runs (13 stimuli × 3 repetitions × 3 runs). After the first practice run, the data were quickly analyzed and feedback was given to the participant on their use of the scale to encourage them as needed to use a wide range of magnitudes in their judgments. Participants were encouraged to use the entire scale so that the stimulus with the highest level of breathiness/roughness presented was rated closer to 1000 and the stimulus with the lowest breathiness/roughness was rated closer to a 1. Participants were told that a rating of 100 is representative as the middle of the scale. Most feedback of this sort took place in the initial loudness training task, but the experimenter reminded the participant of instruction of the task and how to use the scale at the beginning of each task and whenever necessary. The experiment was completed over two sessions approximately 2 hours in duration.

RESULTS

The average of the magnitude estimation judgments from the final two runs (20 total ratings) for each parameter was used in data analysis. Inter-rater and intra-rater correlations were also calculated using Pearson's *r* to assess the reliability of ratings[5,30] between each pair of participants and repetitions of each stimulus as reported below.

Magnitude estimation

Practice stimuli

The dashed lines of Figure 3 show the loudness ME judgments for 25 listeners across a 32-dB range (in 9 steps spaced 4 dB apart). The data are typical for group ME data over a wide range, with a tighter cluster of judgments near the upper end of the continuum and more spread at the lower end of the continuum. The mean ME values across listeners is shown by a solid black function and reveals a nearly linear change in log-magnitude with level in dB.

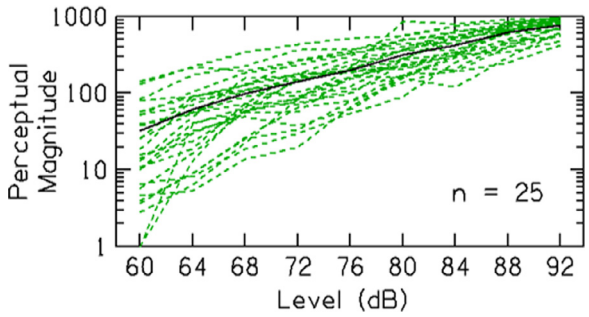

FIGURE 3. Magnitude estimation (ME) of loudness as a function of stimulus level for 25 listeners (dashed lines) and the average across listeners (solid line). The data summarize the practice task completed prior to the breathiness and roughness ME tasks reported below.

Breathy Stimuli: Figure 4 shows the individual data for breathiness ratings for 25 participants as dashed lines; mean data are shown by the solid black line. Judgments were similar across all participants. Breathiness magnitude was highest for the lowest SNR values (the noisiest stimuli) and systematically decreased as the SNR increased. Despite considerable data scatter, the Pearson's correlation coefficient among the 25 participants was high (r = 0.96), showing good inter-rater reliability between all subjects. For intra-rater reliability, the Pearson's correlation coefficient was calculated for the final two judgments of each breathiness condition. The mean intra-rater reliability was 0.96, an indicator that participants were highly consistent within their own judgments for breathiness. There were a few outliers with steeper or shallower slopes than the mean, roughly equally distributed above and below the mean function. The mean function can be taken as the relationship between the physical units (dB SNR) and perceptual magnitude of breathiness.

Rough Stimuli: As for breathiness, Figure 5 shows the individual and mean data for roughness ratings for 25 participants in the same format. The roughness data are similar in

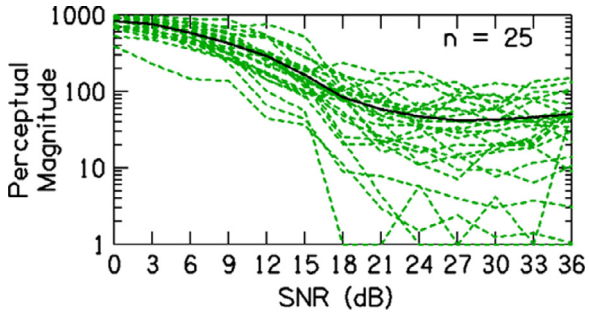

FIGURE 4. Perceived breathiness magnitude as a function of stimulus SNR (dB) for 25 listeners. The solid curve represents the mean across listeners.

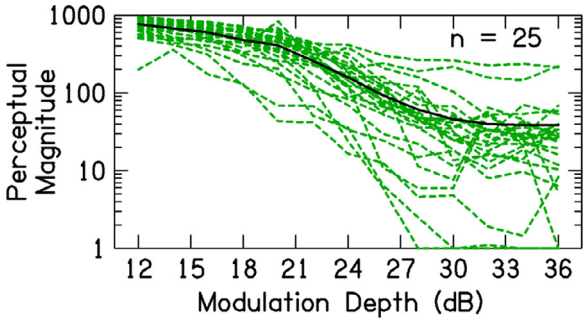

FIGURE 5. Perceived roughness magnitude as a function of stimulus modulation depth (dB) for 25 listeners. The solid curve represents the mean across listeners.

form to the breathiness data (Figure 4). Roughness magnitude was highest for the highest modulation depths (greatest fluctuation) and systematically decreased with decreasing modulation depth. The Pearson correlation coefficient for roughness judgments was 0.94, showing high inter-rater reliability between all subjects. The Pearson's correlation coefficient for intra-rater reliability of 0.94, indicating that participants were highly consistent within their own judgments for roughness as well. Again the mean function can be taken as the relationship between the physical units (dB amplitude modulation depth) and perceptual magnitude of roughness.

Physical scales to psychophysical scales

After initial analysis of both voice quality percepts, raw data were fitted with a logistic function:

$$y = a_0 + (1000 - a_0)./\left(1 + \exp\left(-a_2 * (x - a_3)\right)\right).$$

Where x is the value of independent variable (dB SNR or dB modulation depth), a_0, a_2, and a_3 are coefficients of the logistic equation. The mean data from Figures 4 and 5 are shown in Figure 6 (left and right panels, respectively) along with curves representing the fitted functions. The logistic functions provide a good fit ($r^2 = 0.99$ for both breathy and rough data) to the data over the range of dysphonic voices yet deviate somewhat from the perceptual data at the normal end of the continuum. This is a chronic issue in evaluation of sound quality at the end of the continuum at which the quality diminishes to an imperceptible magnitude (ie, repeatable data are unlikely if one is judging the magnitude of breathiness when essentially no breathiness exists).

Following step three from the methodology of Stevens[17], a single point on the function is arbitrarily chosen as a reference point. For the sone scale, recall that Stevens chose a 1000-Hz tone with a level of 40 dB SPL. For breathiness, we chose as the single reference value the lower x-axis value in the left panel of Figure 6 which approximated the major inflection point in the fitted function. We then rounded that value to the point at which the fitted function crosses a perceptual magnitude of 100. This reference value is approximately equal to 18 dB SNR and is redefined on the upper x axis as 1 breathiness unit. Likewise, for roughness (Figure 6, right panel), the reference point corresponding to a perceptual magnitude of 100 is mapped to the lower x-axis value of −27 dB amplitude modulation depth is redefined to be 1 roughness unit on the upper x axis.

In step four, all other values on the x axis of each plot in Figure 6 are transformed such the relative perceptual magnitude on a ratio scale is expressed relative to the standard reference from step three. This leads to the complete upper x axis range in each panel of Figure 6. In this case, a doubling of breathiness relative to the reference of 1 breathiness unit is 2 breathiness units while a halving of breathiness relative to the reference of 1 breathiness unit is 0.5 breathiness units. The scales are naturally skewed toward dysphonic voices (physical and psychophysical units above the standard values indicated by blue dashed lines in Figure 6). Below these values, the perception of breathiness or roughness is truncated. This is similar to the range of voice quality magnitude among voices, with a very small range among normal voices and a much larger range among dysphonic voices.

DISCUSSION

Voice quality assessments are important tools for reliably indexing voice quality in patients with dysphonic voices. Popular clinical voice assessments consist of rating scales along several quality dimensions. Despite their clinical utility, such ratings do not support comparisons of magnitude, differences in magnitude, or direction of differences in magnitude. For

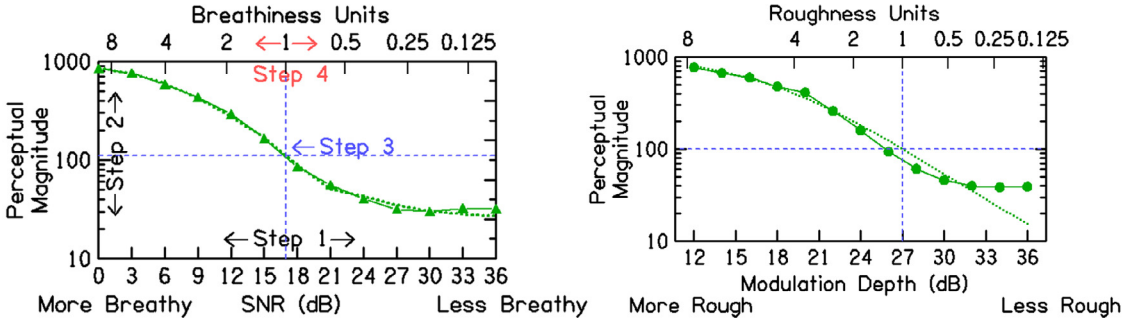

FIGURE 6. Transformation from physical scales (lower x axes) to psychophysical scales (upper x axis) for breathiness (left panel) and roughness (right panel). Steps for scale development, as outlined in the text, are shown in the left panel. The values assigned to 1 scale unit were defined by approximating the major inflection point in the fitted function and rounding to the nearest discrete independent variable value sampled in the corresponding ME experiments.

example, magnitude of change resulting from treatment or disorder progression cannot be expressed based on rating-scale data. Rating scales of voice only produce arbitrary values that may be influenced by contextual bias and often are associated with poor reliability of perceptual judgments. Despite the extensive research on the deficits of these rating scales, little voice quality research has focused on surmounting the limitations associated with rating scales to provide a more concrete way to describe changes in voice quality. The SVMT is one attempt to do so but brings with it another set of challenges, notably the reliance on average data across listeners and the time-consuming nature of the task.

Transforming what is known about the physical units that relate to perceptual breathiness and roughness, based on data from SVMTs, into standard scales can help quantify the change in the perception of voice quality for a single speaker and differences in quality across different speakers. Clinically, standard scales for voice quality may provide more reliable measures in the assessment of dysphonic patients than rating scale measures. In order to do this, physical scales derived from voice quality matching tasks needed to be transformed into psychophysical scales. Psychophysical scales make it possible for changes in voice quality to be described based on magnitude, making it possible to say, for example, that the "The voice quality of a patient improved by 15% after 9 weeks of treatment and by 30% after 12 months of treatment." Thus, physical units and their perceptual equivalent are related through psychophysical scales. Such scales are likely to aid in clinical treatment practice and lead to a greater advancement in the field of voice disorders.

In the present study, listeners were presented with a total of 26 synthetic stimuli (13 breathy and 13 rough) and those listeners evaluating breathiness or roughness magnitude on a magnitude estimation scale from 1 to 1000. Listeners were instructed to use a ratio scale when rating voices in order to understand the perceptual magnitude of roughness and breathiness qualities in order to establish a continuum. As expected, listeners were able to rate both breathiness and roughness stimuli based on the varied SNR (breathiness) and amplitude modulation depth (roughness) levels. Furthermore, the average judgments for each stimulus were similar across listeners. It is expected that there will always be some variability in single judgments when dealing with behavioral measurements, despite the high reliability seen here. For this reason, the average of judgments was used in order to provide a more stable representation of voice quality that could be used to create a psychophysical scale. After fitting raw data from the 25 participants into a function, a reference value was established for breathiness and roughness continuums, similar to that of the sone scale of loudness.

These newly developed scales could be further validated if used with vocal stimuli that have been previously evaluated for breathiness and roughness either through matching tasks or other models that are predictors of voice quality. The applicability of the scales can be improved by assigning names to the scale units analogous to the *sone* unit of loudness. Studying the relationship between the psychophysical scale and already established models of voice quality perception would further demonstrate the practical use of these scales. Quantifying the perception of voice will help generate computational models of voice quality to eventually serve as a more objective evaluation of voice quality to be used in the clinical setting. Future work might also include development of an analogous scale of the strain voice quality.

The use of these scales in both clinical practice and as a tool for research holds great promise as we seek to document, understand, and treat voice disorders. Clinical translation will be implemented via software that may be used to load and play dysphonic voice samples previously collected or to record and play voice samples during an evaluation. This software would comprise synthetic anchor stimuli from the SVMTs with different values of the independent variables spanning the range of perceived breathiness or roughness. Clinicians will be able to listen to several anchor stimuli and assign a scale value on the basis of their perception of the dysphonic voice and anchor stimuli. This software would also incorporate computational models through a click of a button. The computational model output values that correspond to those independent variable values thus can be translated into scale units using the fitted functions in Figure 6. In the current insanitation, breathiness and roughness scaling will be completed separately for the same voice. Future developments on the software will allow voices to be evaluated simultaneously on multiple voice quality dimensions. The results of such a scaling procedure can inform clinical practice in several ways. Perhaps the three most important attributes that can influence clinical practice are: (1) the ability to express magnitude and direction of change (ie, responsiveness to change); (2) the similarity of standard scale value judgments across time, patients, clinicians, and clinics; and (3) the potential to evaluate the same voices with a computational model grounded in the perceptual data and physical stimulus values that were used to develop the scales. The latter could even be used in instances where automated evaluation would be useful or preferred. The translation of scales and the attributes they bring to clinical practice will require practical implementation, validation, potential refinement, and widespread adoption. This process typically unfolds over a several-year period.

CONCLUSIONS

Having developed scales for breathiness and roughness, it is possible to describe breathy and rough voices based on standard units that tell the magnitude and direction of change of each voice quality across treatment sessions. With some listening experience, such units become meaningful and can be used to evaluate the breathiness or roughness of any voice without having any intuitive knowledge of the actual physical units (eg, dB SNR or dB amplitude modulation depth) underlying scale development. As such, these scales provide well defined measurement units with reference values of known scale properties that can be treated using logical and lawful mathematical operations. The ability to quantify the

magnitude of change in a voice is critical to understanding the perception of voice quality in patients with voice changes due to an underlying disorder.

REFERENCES

1. Hirano M. *Clinical Examination of Voice*. New York: Springer-Verlag, Wien; 1981.
2. Kempster GB, Gerratt BR, Abbott KV, et al. Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol. *Am J Speech-Lang Pat*. 2009;18:124–132.
3. Nagle KF. Emerging scientist: challenges to CAPE-V as a standard. *Perspect ASHA Spec Interest Groups*. 2016;1:47–53.
4. Oates J. Auditory-perceptual evaluation of disordered voice quality: pros, cons and future directions. *Folia Phoniatr Et Logop*. 2009;61:49–56.
5. Patel S, Shrivastav R, Eddins DA. Perceptual distances of breathy voice quality: a comparison of psychophysical methods. *J Voice*. 2010;24:168–177.
6. Nemr K, Simões-Zenari M, Cordeiro GF, et al. GRBAS and Cape-V scales: high reliability and consensus when applied at different times. *J Voice*. 2012;26:812.e17–812.e22.
7. Zraick RI, Kempster GB, Connor NP, et al. Establishing validity of the consensus auditory-perceptual evaluation of voice (CAPE-V). *Am J Speech Lang Pat*. 2011;20:14–22.
8. Johnson K, Brehm SB, Weinrich B, et al. Comparison of the pediatric voice handicap index with perceptual voice analysis in pediatric patients with vocal fold lesions. *Arch Otolaryngol Head Neck Surg*. 2011;137:1258–1262.
9. Karnell MP, Melton SD, Childes JM, et al. Reliability of clinician-based (GRBAS and CAPE-V) and patient-based (V-RQOL and IPVI) documentation of voice disorders. *J Voice*. 2007;21:576–590.
10. Kreiman J, Gerratt BR, Kempster GB, et al. Perceptual evaluation of voice quality: review, tutorial, and a framework for future research. *J Speech Lang Hear Res*. 1993;36:21–40.
11. Kelchner LN, Brehm SB, Weinrich B, et al. Perceptual evaluation of severe pediatric voice disorders: rater reliability using the consensus auditory perceptual evaluation of voice. *J Voice*. 2010;24:441–449.
12. Solomon NP, Helou LB, Stojadinovic A. Clinical versus laboratory ratings of voice using the CAPE-V. *J Voice*. 2011;25:e7–e14.
13. Chan KM, Yiu EM. The effect of anchors and training on the reliability of perceptual voice evaluation. *J Speech Lang Hear Res*. 2002;45:111–126.
14. Eadie TL, Baylor CR. The effect of perceptual training on inexperienced listeners' judgments of dysphonic voice. *J Voice*. 2006;20:527–544.
15. Eadie TL, Kapsner-Smith M. The effect of listener experience and anchors on judgments of dysphonia. *J Speech Lang Hear Res*. 2011;54:430–447.
16. Shrivastav R, Sapienza CM, Nandur V. Application of psychometric theory to the measurement of voice quality using rating scales. *J Speech Lang Hear Res*. 2005;48:323–335.
17. Stevens SS. A scale for the measurement of a psychological magnitude: loudness. *Psychol Rev*. 1936;43:405–416.
18. Eddins DA, Shrivastav R. Psychometric functions for rough voice quality. *J Acoust Soc Am*. 2010;127. 2021-2021.
19. Gelfand SA. *Hearing: An Introduction to Psychological and Physiological Acoustics*. New York: Marcel Dekker, Inc; 1998:262–263.
20. Stevens SS. On the theory of scales of measurement. *Science*. 1946;103:677–680.
21. Stevens SS. Ratio scales, partition scales, and confusion scales. In: Gulliksen H, Messick S, eds. *Psychological Scaling: Theory and Applications*. New York: Wiley; 1960.
22. Stevens SS. The psychophysics of sensory function. In: Rosenblith WA, ed. *Sensory Communication*. Cambridge Mass: MIT Press; 1961.
23. ANSI. *ANSI Sl.1-2010, American National Standard Acoustical Terminology*. New York: American National Standard Institute; 2010.
24. Patel S, Shrivastav R, Eddins DA. Developing a single comparison stimulus for matching breathy voice quality. *J Speech Lang Hear Res*. 2012;55:639–647.
25. Patel S, Shrivastav R, Eddins DA. Identifying a comparison for matching rough voice quality. *J Speech Lang Hear Res*. 2012;55:1407–1422.
26. Fastl H. Roughness and temporal masking patterns of sinusoidally amplitude modulated broadband noise. In: Evans EF, Wilson JP, eds. *Psychophysics and Physiology of Hearing*. London: Academic; 1977:403–414.
27. Fastl H, Zwicker E. *Psychoacoustics: Facts and models*. 3rd ed New York: Springer; 2007.
28. Eddins DA, Shrivastav R. Psychometric properties associated with perceived vocal roughness using a matching task. *J Acoust Soc Am*. 2013;134:EL294–EL300.
29. Kemp S. Roughness of frequency-modulated tones. *Acta Acust United Ac*. 1982;50:126–133.
30. Shrivastav R, Camacho A, Patel S, Eddins DA. A model for the prediction of breathiness in vowels. *J Acoust Soc Am*. 2011;129:1605–1615.