Published in final edited form as:

J Speech Lang Hear Res. 2012 October ; 55(5): 1407–1422. doi:10.1044/1092-4388(2012/11-0160).

Identifying a Comparison for Matching Rough Voice Quality

Sona Patel^a, Rahul Shrivastav^{a,b}, and David A. Eddins^c

^aUniversity of Florida, Gainesville, FL

^bMalcom Randall VAMC, Gainesville, FL

^cGlobal Center for Hearing and Speech Research, University of South Florida, Tampa, FL

Abstract

Purpose—Perceptual estimates of voice quality obtained using rating scales are subject to contextual biases that influence how individuals assign numbers to estimate the magnitude of vocal quality. Because rating scales are commonly used in clinical settings, assessments of voice quality are also subject to the limitations of these scales. Instead, a matching task can be used to obtain objective measures of voice quality, thereby facilitating model development and tools for clinical use.

Method—Twenty-seven individuals participated in a rating task or at least 1 of 3 matching tests (named after their modulation functions: SINE, SQUARE, POWER) to quantify the degree of roughness in dysphonic voice stimuli. Participants evaluated the roughness of 34 voice samples using an amplitude-modulated complex carrier.

Results—The matching thresholds were highly correlated with the ratings estimates. Reliability of thresholds did not significantly differ across tasks, but linear regressions showed that the POWER test resulted in larger perceptual distances.

Conclusions—A matching task can be used to obtain reliable estimates of roughness in dysphonic voices. The POWER comparison is recommended because the variability in matching thresholds across the range of roughness was evenly distributed, and the perceptual distances between stimuli were maximized.

Keywords

voice quality; roughness; rating scale; matching; amplitude modulation

Roughness is an important dimension of dysphonic voice quality and may be defined as the psychoacoustic impression of irregularity of vocal fold vibrations (Hirano, 1981). Perceptual evaluation of roughness is a critical part of many research studies and clinical protocols (e.g., the Consensus Auditory-Perceptual Evaluation of Voice or CAPE-V, Kempster, Gerratt, Verdolini-Abbott, Barkmeier-Kraemer, & Hillman, 2009; and the Grade, Roughness, Breathiness, Asthenia, Strain or GRBAS, Hirano, 1981). The current experiments are part of a larger series of ongoing studies designed to quantify the perception of rough voice quality and to develop computational models of roughness perception that may both replicate listener perception of voice quality and provide predictions that can be used in clinical activities such as for assessment or treatment. As such, the goals of this experiment include development and evaluation of psychophysical methods and stimuli that lead to valid, reliable, repeatable, and concise estimates of roughness perception and that produce data

that are suitable for computational modeling and detailed acoustic analyses. As in our previous study on breathy voice quality (Patel, Shrivastav, & Eddins, 2010, 2012), the approach taken here includes the use of a psychophysical matching task to minimize context and bias effects. Similar to that study, the initial work reported here focuses on the development of a single comparison stimulus for the matching task that is suitable for voices that span the spectrum of roughness.

Issues With Rating Scale and Magnitude Estimation Tasks

Research targeting rough voice quality per se is limited. Most have used some form of a rating scale task (e.g., an n -point rating scale or a visual-analog scale; Deal & Emanuel, 1978; de Krom, 1995; Martin, Fitch, & Wolfe, 1995) or a direct magnitude estimation task (Toner & Emanuel, 1989) to estimate the roughness of a set of stimuli. Although these tasks are very simple to conduct, the numbers that listeners assign to the percepts are arbitrary and biased by the presentation context. By *arbitrary*, we mean that the raw values assigned to a particular stimulus do not always represent the same perceptual magnitudes. For example, the value 3 on a 7-point rating scale for roughness may refer to a different magnitude of roughness across listeners, stimulus sets, or experiments. Unlike other measurements, such as physical measures of length, weight, or time, rating scale values are not made with reference to a universally accepted standard. Thus, these values are subject to contextual biases and tend to be highly variable across listeners and experiments. By *contextual biases*, we mean that the raw value assigned to a stimulus on a rating scale or magnitude estimation task is affected by factors other than the magnitude of the percept itself. For example, the scores on such tasks are influenced by the range of an attribute present in a stimulus set (“range effects”), by the frequency with which specific stimulus attributes are repeated within an experiment (“frequency effects”), or even by where the stimuli lie along a perceptual continuum of magnitude (“bow effects”; e.g., see Guilford, 1954; Parducci & Wedell, 1986; Poulton, 1989; Siegel, 1972; Stewart, Brown, & Chater, 2005). Indeed, some of these biases likely result from the arbitrary nature of the ratings assigned to stimuli.

Another related problem with rating scale tasks—particularly as used in most clinical applications—is that these only provide ordinal data (i.e., rank order across stimuli) rather than interval or ratio measurements. Interval-level measurements require that successive units on the measurement scale have equal distances, whereas ratio scales require successive units to have equal ratios (see Stevens, 1946, for more details on measurement scales). Because the values assigned on rating scales are arbitrary and context-dependent, it is not possible to determine the underlying relationship between successive items in a rating scale. In some cases, when certain assumptions about the distribution of the perceptual data can be met, it may be possible to determine interval-level measurements from rating scale data (see “Law of Comparative Judgment,” Thurstone, 1927; also see Guilford, 1954; Shrivastav, Sapienza, & Nandur, 2005). However, practical limitations during clinical use of rating scales, such as the inability to obtain multiple judgments or statistical normalization, often preclude such secondary analyses of rating scale data. By definition, perceptual data on an interval scale must be such that items identified as being equidistant on the rating scale must also be perceptually equidistant. Suppose we have three stimuli that are judged as 2, 3, and 4 on a 7-point rating scale of roughness, respectively. For these data to be on an interval scale, the difference in perceived roughness (not just the numeric scores) between stimuli judged to be 2 and 3 must be equal to the difference in roughness between stimuli judged to be 3 and 4. Unfortunately, there is no evidence that such interval relationships hold when listeners judge voice quality (or most other psychophysical continua). Most experiments or clinical protocols on voice quality do not instruct listeners to attempt using the rating scales with fixed perceptual intervals. Furthermore, the arbitrary use of numbers and the high context dependencies further limit the confidence in how listeners may be using the scale

values. For these reasons, it has been recommended that rating scale data, particularly the way ratings are elicited in a clinic, are best treated as ordinal data (Shrivastav et al., 2005).

These factors severely limit the utility of perceptual data obtained using rating scale or direct magnitude estimation tasks for the development of computational models, as well as for clinical use itself. For example, the contextual biases introduce significant variability in perceptual data that is difficult to explain by computational models that only use stimulus acoustic characteristics as predictor variables. The ordinal nature of the data allows clinicians and scientists to ascertain whether the magnitude of a particular voice quality dimension is greater or less across two or more stimuli, but does not allow judgments that can state the degree to which one stimulus differs from another. Thus, although clinicians can ascertain that a patient's voice improved, they cannot state the magnitude of improvement with any confidence.

The Matching Task

Instead of rating scale or direct magnitude estimation, a matching task can be used to obtain perceptual judgments of voice quality. This task avoids errors resulting from arbitrary assignment of numbers, minimizes contextual biases, and allows perceptual measurement with ratio-level properties. In this task, the magnitude of a quality is measured by comparing the test stimulus (the "standard") to a single reference stimulus (the "comparison"). However, instead of simply assigning the standard stimulus a numeric score, participants manipulate a predefined parameter of the comparison until the quality of the comparison matches that of the standard. A physical value describing the acoustic characteristics of the comparison stimulus at the point of subjective equality between the two stimuli is then used as a measure of the test stimulus's voice quality. By avoiding arbitrary assignment of numbers to denote the magnitude of a percept, the contextual biases are minimized and greater reliability in perceptual judgments is obtained (Kreiman, Gerratt, & Ito, 2007; Patel et al., 2010, 2012). Furthermore, by virtue of comparing stimuli to a fixed or constant reference stimulus, it also becomes possible to compare data across multiple experiments or participants (Patel et al., 2010, 2012). The matching task is commonly used in all aspects of sensation and perception, including vision, audition, taste, smell, and touch (Belkin, Martin, Kemp, & Gilbert, 1997; Stevens & Hall, 1966). In the broader context of psychophysics, it is generally preferred over ratings, direct magnitude estimation, and magnitude production tasks because it avoids many of the biases and pitfalls associated with those tasks.

Stimulus continua may vary along one dimension or more than one dimension. In perception, however, it is commonplace to use a comparison stimulus that varies along one dimension to evaluate a stimulus of interest that varies along multiple dimensions. This is precisely the purpose of the matching experiment that has been used extensively in psychophysics (in audition, vision, touch, taste, and smell) and adopted here. Its reliability and validity are the reasons that matching was the method of choice for designing standard scales of measurement, such as the sone scale, and those measurements are so widely recognized that they are the bases for international standards for perceptual judgments (e.g., ANSI S3.4-2007). Furthermore, there is a large body of evidence indicating that subjects are quite good at doing this sort of comparison. A classic example is the perception of loudness, where the comparison stimulus varies only along one physical dimension (intensity), and stimuli to be evaluated vary along many perceptual dimensions (i.e., loudness, pitch, roughness, sharpness, timbre). Consider judging the loudness of the sounds coming from the small loudspeakers connected to the computer. The loudness of these sounds can be judged reliably even though as the intensity is increased from very low to very high, a broader bandwidth is perceived, various distortions become audible, and at the upper end those distortions become prominent. Thus, human observers are quite good at judging a single

percept in the presence of multidimensional stimuli characterized by multiple perceptual attributes.

Although it is possible to overcome some of the biases of rating scales by averaging and/or standardizing multiple judgments of the same stimulus (e.g., “random” and “criterion bias”; see Shrivastav et al., 2005), other limitations, such as range and frequency effects, cannot be easily overcome. The goal of developing the matching task described here is to enable voice quality measurements with greater precision and with known measurement properties. This is essential for the development of computational models of voice quality perception. These models will have direct clinical utility because they can provide a set of tools to quantify changes in the magnitude of voice quality accurately, efficiently, and in a standardized manner.

The matching task described here could be used in everyday clinical practice; nevertheless, it has not been used as such, primarily because of the time required to obtain perceptual data. The procedures described here are intended for use in the laboratory, where higher precision in measurement is of greater importance than the time required to obtain measurements. There is a tradeoff in the accuracy of measurements, the complexity of the psychophysical task employed, and the time required for measurement (see Shrivastav, 2011, for a review of some psychophysical measurement approaches). It is our goal to use this higher precision psychophysical method to help develop computational models of voice quality perception. These models, in turn, can be used for fast and accurate measurement of voice quality in everyday clinical practice. Even so, if high accuracy, context independence, and ratio-level measurement are desired, the approach described here can be used in clinical practice as well. An abbreviated matching task based on standard voice quality scales could be developed in the future and could be clinically viable. Importantly, the use of physical measurement scales based on the data obtained in large-scale matching tasks like the one reported here may be used to obtain objective values that serve as the basis for modeling voice quality (e.g., see Shrivastav, Camacho, Patel, & Eddins, 2011), and such models may eventually lead to an automatic method for quantifying disordered vocal qualities in the clinic.

A somewhat different matching approach to quantify dysphonic voice quality has also been described by Gerratt and Kreiman (2001), who used a custom-designed speech synthesizer to create a synthetic copy of a dysphonic voice stimulus. In this approach, listeners were allowed to manipulate multiple synthesizer parameters until the synthesizer output was judged to be an exact match to the test stimulus. The values of the various synthesizer parameters at the point of subjective equality were used to describe the quality of the stimulus.

Unfortunately, the data obtained from the multi-parameter matching approach do not lend themselves to the development of computational models of voice quality. This is because it remains unclear whether a given multiparametric description of voice quality reflects a single perceptual magnitude or whether multiple parameter settings interact in some way to reflect the same identical quality. Note that such trading cues are common in psychoacoustics and in speech perception. For example, loudness of a sound can be affected not only by the sound pressure level, but also by its spectral slope. Similarly, the voicing characteristics for a stop sound can be cued not just by the voice onset time, but also from the burst duration, the overall duration, the spectral tilt, and even the duration of the preceding vowel in a syllable (Kewley-Port, Pisoni, & Studdert-Kennedy, 1983). Furthermore, the perceptual distances between different stimuli measured using a multiparametric matching approach cannot be easily ascertained. Thus, it is difficult to judge

from the multiparametric matching approach how two stimuli vary along one or more perceptual dimensions and what the magnitude of these differences might be.

In contrast, the single-variable matching task is one of the classic methods used in the study of sensation and perception and is commonly used to scale perceptual magnitudes for stimuli that evoke multidimensional perceptual attributes (Stevens, 1975). Such tasks have been used extensively in all sensory domains, including psychoacoustic scaling of loudness and pitch (Stevens & Hall, 1966; Stevens & Volkman, 1940). In comparison to the multiparameter matching task, the single-variable matching task is simpler to perform and the results are more straightforward to interpret. For example, it is easier to adjust the intensity of a pure tone to closely match the perceived loudness of a complex tone than it is to adjust the intensity, fundamental frequency, spectral slope, and component phases of a complex tone to simultaneously match the loudness, pitch, timbre, roughness, and sharpness of that complex tone. Another advantage of the single-variable matching task is its ease of quantification, which simplifies the estimation of perceptual distances between stimuli, while the physical units that correspond to subjective magnitude facilitate the modeling of perceptual voice quality attributes.

Goals

The matching approach described here is based on Patel et al. (2012), who demonstrated that a single-variable matching task resulted in sensitive and reliable measurements of breathy voice quality. In that study, participants were required to manipulate a single attribute (the signal-to-noise ratio [SNR]) of a synthetically generated comparison (a low-pass filtered sawtooth wave mixed with speech-shaped noise) until the comparison and standard were perceived to have equal breathiness. The SNR of the comparison at the point of subjective equality was then used as a measure of breathiness. In the present study, the single-variable matching task described by Patel et al. (2012) was modified to evaluate the perception of rough voice quality in vowels. Roughness was measured by comparing a set of dysphonic voice stimuli (“standards”) with a reference stimulus (“comparison”). The comparison stimulus used here also consisted of a sawtooth wave mixed with speech-shaped noise. However, the SNR was maintained at a constant level, and a different variable parameter—amplitude modulation depth—was manipulated to alter the roughness of the comparison. This variable parameter was selected because prior research has shown that roughness of sinusoidal stimuli increased monotonically with an increase in the depth of their amplitude modulation (Zwicker & Fastl, 1990). The sawtooth-and-noise carrier signal has a quality consistent with vowel sounds while it avoids the tonal quality of sine-wave stimuli. The frequency of amplitude modulation was held constant. The modulation depth at which the comparison and standard were perceived to have equal roughness was used as an index of roughness.

Four related experiments are described. Experiment 1 was a rating task and was performed as a baseline measure of rough voice quality to which the single-variable matching task could be compared and contrasted. Three different adaptations of the matching task protocol were sequentially designed and tested. These varied in terms of the complexity of the modulation function, with some additional differences in the familiarization exercises, in an attempt to obtain an adequate range of perceived roughness and matching thresholds with minimal variability that are consistent with previous roughness data. The first matching experiment (Experiment 2) used a simple sine-wave modulator (henceforth referred to as the SINE experiment). The second and third matching experiments (Experiments 3 and 4) included increasingly complex modulation functions—a square-wave and a raised-sinusoidal function, respectively (referred to as the SQUARE and POWER experiments). The principal distinguishing feature among these three modulators is their temporal

envelope shapes, which in turn dictate the amount of perceived fluctuation and quantitative fluctuation. *Fluctuation* is commonly quantified as the fourth moment of the waveform or the waveform crest factor (peak divided by the root-mean-squared amplitude). In the absence of prior roughness matching data, this range of modulators was chosen a priori as one that likely would encompass the perceived range of roughness observed in dysphonic voice stimuli.

In addition to modifications of the amplitude modulation function, the experiments used familiarization exercises along with each of the matching tasks. The use of training exercises may aid participants in completing a novel task such as the matching task and can be used as a means of improving the reliability of rating perceptual phenomena (Martin & Wolfe, 1996). Participants performed one or more familiarization exercises prior to performing the main matching test. These included a matching familiarization (MF) task, an interactive quality demonstration (IQD) task, and a visual sort ranking (VSR) task. The number of training exercises and the task instructions given to an individual depended upon the experiment. All training was performed prior to beginning the main matching test. Participants in the POWER experiment were trained using the VSR prior to each session.

The matching thresholds obtained using each of these three comparison stimuli were compared to rating scale estimates (Experiment 1) because these are the most commonly used judgments and are considered by some researchers to be the standard for quantifying changes in roughness. As described previously, although a good correlation between the rating scale data and matching threshold would indicate consistency in measurement, the absolute values obtained in the rating scale task would vary from one experiment to the next, whereas the matching thresholds are expected to remain relatively constant.

Experiment 1: Rating Scale Measure of Roughness

This preliminary experiment was conducted to obtain ratings of roughness using a 5-point rating scale similar to the paradigm of Deal and Emanuel (1978), except that the rating for a stimulus was computed as the mean across participants. Also, participant ratings were measured as the mean rating across repetitions. These data served as a reference for comparison to previous data in the literature and to the data obtained from the novel matching experiments that follow. All experiments described here were approved by the University of Florida Institutional Review Board and all subjects consented to participate. Participants were paid \$5/hr for completing this study and received a \$10 bonus upon completing all test sessions for any experiment that required more than one test session.

Method

Participants—Five individuals, all graduate students in speech-language pathology at the University of Florida, participated in this experiment (three women, two men). Listeners' ages ranged from 21 to 40 years. Participants had normal hearing bilaterally, confirmed through a hearing screening (air-conduction pure-tone threshold below 20 dB HL at 250 Hz, 500 Hz, 800 Hz, 1000 Hz, 2000 Hz, and 4000 Hz; ANSI S3.21-2004).

Standard stimuli—Thirty-four samples of a sustained /a/ vowel were evaluated in this experiment. These were natural voice samples selected from the Kay Elemetrics Disordered Voice Database (KEDVD; Kay Elemetrics, Inc., Lincoln Park, NJ) to represent a wide range of roughness (from nearly normal to severe roughness). The /a/ vowel was chosen because it can be modeled as an all-pole filter, assuming a linear source-filter model of speech production, with the formants separated from the fundamental frequency. The stimuli were edited to include only a 500-ms segment from the middle of the waveform. A 10-ms cosine onset and offset ramp were applied to all stimuli to avoid artifacts during playback.

Following the nomenclature used in many psychoacoustic experiments, these test stimuli are referred to as the “standards” in each of the three matching experiments that follow.

Equipment—All tests were conducted in a single-walled sound-treated booth. Stimulus generation, presentation, and response collection for this perceptual test (and the training and listening tests that follow) were controlled by the TDT SykofizX software application (Tucker-Davis Technologies, Inc., Alachua, FL) using a TDT RP2 real-time processor, TDT PA5 programmable attenuators, and a TDT HB7 headphone buffer. The stimuli were delivered to the right ear of the subject via high-fidelity ear inserts (ER2, Etymotic Research Inc., Elk Grove Village, IL) designed to deliver a flat frequency response at the ear drum from 100 to 10000 Hz. Monaural presentation was chosen because modeling of voice quality often depends on loudness models that do not easily account for binaural interaction effects. The output level was calibrated to ensure that each stimulus was delivered at 75 dB SPL.

Procedure—In this experiment, each trial consisted of a single standard stimulus. Participants rated the roughness of each standard using a scale from 1 to 5 in steps of 1, where 1 represented *minimum roughness* and 5 represented *severe roughness*. Responses were made by clicking the appropriate button on a computer monitor. Each of the 34 standards was presented 10 times in random order.

Results

The ratings averaged across participants are shown in Figure 1, with error bars representing the standard error of the mean. The ratings are presented for the 34 standard stimuli ordered from most to least roughness. In rating the roughness of these voice samples, listeners used the full 5-point scale, as indicated by the dispersion of data along the vertical dimension. In general, the variability among subjects was highest for items in the middle of the scale and lowest at either end. To investigate whether the variability for a particular standard was due to outliers, boxplots of the voice standards were computed for each standard (see Figure 2). The interquartile ranges (IQRs; represented by the boxes) indicate the range of listener scores between the 25th and 75th percentiles. This figure confirms the greater variability seen in the middle of the scale. Figure 3 shows the IQRs regressed onto the listener ratings. Nonlinear (polynomial) regression provided a better fit of the data than linear regression ($R^2 = .456$ and $.169$, respectively). This is likely a manifestation of the “bow effect” that is common to rating scales across a wide range of psychophysical continua.

Before performing further statistical analysis, a one-sample Kolmogorov-Smirnov test was performed for each of the 34 standards to determine whether the listener ratings were normally distributed around the mean (at significance level $p = .05$). The null hypothesis is that each variable has a standard Gaussian distribution. None of the results was significant ($p > .05$), suggesting that all variables demonstrated Gaussian behavior and supporting the use of linear statistics. Then, participant reliability was determined as the mean Pearson’s correlation coefficient between and within listeners. Mean interjudge reliability was computed as the mean of a series of correlations between each participant’s mean ratings and the remaining participants’ mean ratings. For each correlation, each of the 34 values in one vector was computed as the mean estimate across the 10 assessments from each listener and in the other vector, the mean estimate across the remaining participants’ mean ratings. The mean interjudge reliability or measure of between-subjects reliability was 0.90 ($SD = 0.03$, $M = 0.90$). Intrajudge reliability was calculated as the mean of a series of correlations among the 10 judgments within a participant. The mean intrajudge reliability score across all participants or measure of within-subject reliability was 0.82 ($SD = 0.07$, $M = 0.84$). Together, these results indicate that average participant ratings were reliable with each other,

but that individual participants were not as reliable among their repeated judgments. These values are better than those reported by Kreiman, Gerratt, and Berke (1994), in which ratings of roughness were obtained using a 7-point rating scale from eight listeners. The reliability was reported using Pearson's correlations with intrarater correlations from 0.66 to 0.91 ($M = 0.78$) and interrater correlations from 0.22 to 0.82 ($M = 0.60$).

Experiment 2: Roughness Matching Using SINE Modulation

The primary goal of this experiment was to establish the feasibility of using a psychophysical matching task to quantify rough voice quality. A secondary goal was to evaluate the suitability of a comparison stimulus that involved the superposition of sinusoidal amplitude modulation on the carrier stimulus used previously to quantify breathiness (Patel et al., 2012). In contrast to the matching task of Patel et al. (2012), in which the SNR of the comparison stimuli was varied to match the perceived breathiness of the standard stimuli, the depth of amplitude modulation of the comparison stimulus was varied to match the perceived roughness of the standard stimuli.

Method

Participants—Ten individuals (nine women, one man; age range = 19–23 years, $M = 21$ years) were recruited to participate in this experiment. Participant inclusion criteria and payment schedule were identical to Experiment 1.

Comparison stimulus—Perceived roughness is often measured in terms of the depth of sinusoidal amplitude modulation applied to a carrier stimulus (e.g., Zwicker & Fastl, 1990). Thus, the comparison stimulus used here was composed of a carrier that was amplitude modulated. The carrier was a sawtooth wave with a fundamental frequency of 151 Hz mixed with broadband noise that was low-pass filtered above 151 Hz at -7.262 dB/octave. The characteristics of the carrier stimulus were selected as the average of all voices within the KEDVD and are identical to the comparison stimulus used by Patel et al. (2012) for matching breathiness. The SNR of the comparison was fixed at 12 dB. The waveform of the sawtooth-and-noise complex was then amplitude modulated by a $\sin$ wave as shown in Equation 1:

$$Y(t) = 1 + m[\sin(2\pi ft - \theta)] * c(t), \quad (1)$$

where m is modulation depth between a minimum of 0 and a maximum of 1, f is modulation frequency in Hz, t is time in seconds, θ is starting phase, and c is the sawtooth-and-noise carrier. In this experiment, the modulation frequency f was fixed at 25 Hz, and the modulation phase was fixed at 0 radians. The modulation depth was the independent variable. By convention in auditory perceptual experiments, the modulation depth or modulation index was varied on a logarithmic scale as shown in Equation 2:

$$\text{Modulation_index(dB)} = 20 * \log_{10}(m). \quad (2)$$

A sample of the sinusoidal modulation waveform (without a carrier) is shown in Figure 4 for a modulation depth of 0 dB or $m = 1$.

Procedure—Before beginning the matching test, participants were provided one training exercise, referred to as the matching familiarization (MF) task. The purpose of the MF task was to help participants become accustomed to the matching interface and stimuli. The MF task was identical to the listening test described below, except that it included a limited set of six standards that were not included in the main listening test (also selected from the

KEDVD). Although no feedback was provided, participants were allowed to repeat the task as often as needed to familiarize themselves with the experimental procedure. Approximately 30 min were needed to complete this task.

Once participants had completed the initial exercise, they began the main matching test. The stimulus set included the same 34 standards described in Experiment 1. In this experiment, a pair of stimuli was presented on each trial. The first stimulus in each pair was the standard being evaluated, and the second stimulus was the comparison stimulus. The participant was asked to judge whether the roughness of the comparison was greater or less than the roughness of the standard by clicking the appropriate button on a computer monitor. If the comparison was perceived to be less rough than the standard, then the amplitude modulation depth (in dB) of the comparison stimulus was increased for the next trial. This resulted in greater roughness in the subsequent comparison stimulus. On the other hand, if the comparison was perceived to have greater roughness than the standard, its amplitude modulation depth was decreased for the next trial. When the participant perceived the roughness in the comparison to be equal to the roughness in the standard, he or she was instructed to click on a button marked “equal roughness.” To ensure that a range of independent variable values was explored on each condition, the “equal” button was not enabled until the fourth reversal in direction of the independent variable value (i.e., increasing and decreasing modulation depth). The modulation depth of the comparison at this subjective point of equality was termed the *matching threshold* and used as an index of the roughness for that standard stimulus.

The roughness matching task was completed at least six times (i.e., six runs) for each standard using two different initial values of modulation depth. For three of the runs, the initial modulation depth was set to a very low value, so participants initially were forced to increase the roughness of the comparison. These are referred to as “ascending runs,” as the value of the modulation depth on successive trials increased or ascended from the initial modulation depth (Gescheider, 1976). For the other three runs, the initial modulation depth was set to a very high modulation depth, resulting in “descending runs.” Thresholds from the ascending and descending runs were averaged to obtain a single matching threshold for each standard and from each participant. This test required 5 to 8 hr to complete for all 34 standards. Participants were tested in three to five sessions of approximately 1.5 to 2 hr each. The hardware, software, and test location were the same as in Experiment 1.

Results

The mean matching thresholds for the SINE task are shown in Figure 5 (top panel), with modulation depth (in dB) on the ordinate and the 34 standard stimuli on the abscissa ordered from highest to lowest matched threshold. The standards are labeled according to the numerical labels assigned in the rating task to maintain labeling consistency across experiments. Matching thresholds in units of modulation depth ranged from –21.8 dB (least roughness) to –8.1 dB (most roughness) with a mean threshold of –16.2 dB across standards. The small error bars indicate that listeners were able to consistently differentiate levels of roughness across samples.

A one-sample Kolmogorov-Smirnov test was performed for each of the 34 standards to determine whether the listener ratings were normally distributed around the mean. None of the results was significant ($p > .05$). Therefore, participant reliability was computed using a procedure similar to that of Experiment 1. Here, interjudge reliability was calculated as the Pearson’s correlation coefficient among the 10 mean judgments across participants. Intrajudge reliability was computed as the Pearson’s correlation coefficient among the six judgments within a participant. Results for interjudge reliability were modest, with a mean

of 0.67 ($SD = 0.05$). In contrast, the mean intrajudge reliability was 0.90 ($SD = 0.07$), suggesting that participants were highly consistent in their own judgments.

To examine the overall dispersion in the data, box-plots of the 34 standards were examined (shown in Figure 6, top panel). The standards are in order of high to low roughness, identical to Figure 5 to aid in comparisons. This figure reveals larger IQRs for standards with greater roughness (i.e., Standards 1, 3, 4, etc.). To further investigate this trend, the IQRs for each standard were regressed onto the listener thresholds (shown in Figure 7, blue solid line and asterisks). The dashed and dotted lines and other symbols represent data from Experiments 3 and 4 below and will be discussed later. The linear regression revealed a positive trend, with a slope of 0.157 ($R^2 = .12$, $p = .043$), indicating a greater variability in thresholds with more roughness (small absolute values of modulation depth). The higher variability associated with the roughest standards likely reflects greater difficulty matching the roughness of these standards to the comparison. This is because even at the maximum modulation depth, the sinusoidal modulation function did not result in the same magnitude of roughness as observed in some of the dysphonic voices.

The thresholds obtained in the SINE matching task were then compared to rating scale judgments to determine whether the matching task resulted in similar perceptual distances as the rating scale procedure. The SINE thresholds were regressed onto the ratings in the left panel of Figure 8. Linear regression revealed a positive relationship that accounted for 86.7% of the variance in the data ($p < .001$), suggesting a strong relationship between the two measures of roughness. The Pearson's correlation coefficient between ratings and SINE thresholds was high and statistically significant ($r = .93$, $p < .01$).

These results indicate that the average matching thresholds may be a good measure of roughness. In addition, unlike rating scales, roughness can be measured with a specific unit (e.g., modulation depth measured in dB) that can be used consistently across multiple experiments. However, the sine-wave modulation resulted in high variability in matching thresholds, particularly for the severely rough standards.

Experiment 3: Roughness Matching Using SQUARE Modulation

In this experiment, the modulator used in creating the comparison stimulus had a square-wave envelope rather than the sine-wave envelope of Experiment 2. In addition, a brief familiarization period as well as an interactive voice quality demonstration were introduced before the test.

Method

Participants—Five individuals (two women, three men; age range = 19–51 years, $M = 30.6$ years) took part in the SQUARE experiment. Participant inclusion criteria and payment schedule were identical to Experiment 1.

Comparison stimulus—The comparison stimulus used for this experiment was identical to that used in the SINE experiment, except that the modulation function was modified to increase the maximum roughness of the comparison. Zwicker and Fastl (1990) showed that square-wave amplitude modulation resulted in greater roughness than sine-wave modulation for the same carrier waveform. Thus, the modulation function was modified to be a 40-Hz square wave as shown in Equation 3:

$$Y(t) = 1 + m \left[\sum_{n=1}^N \frac{1}{n} * \sin(2\pi f_n t + \theta) \right] * c(t), \quad (3)$$

where $n = 1, 3, 5, \dots, 39$. A segment of the modulating waveform is shown in Figure 4.

Procedure—Individuals who participated in this experiment performed two familiarization exercises prior to the main matching test. An “interactive quality demonstration” (IQD) was provided to help differentiate between variations in roughness and variations in other percepts such as pitch and breathiness. This is critical because dysphonic voices often co-vary across multiple voice quality dimensions, and listeners often attend to global changes in voice quality rather than focus upon changes in one specific dimension of voice quality. It was assumed that providing stimuli that varied substantially in roughness, breathiness, and pitch would ensure that participants attended primarily to the roughness dimension during the main test without explicitly defining roughness. The IQD was developed as a graphical user interface (GUI) in MATLAB (Math-Works, Inc., Natick, MA) and contained three series of voice stimuli. The first series consisted of synthetic stimuli differing in two levels of pitch (low and high) generated by altering the fundamental frequency of the synthetic stimulus. The second series consisted of three synthetic stimuli varying in breathiness (marked as high, low, and none). These were generated by changing the aspiration noise level (AH) and open quotient (OQ) in a sample generated using the Klatt synthesizer (HLSyn; Sensimetrics Corp., Malden, MA). The third series consisted of three stimuli varying in roughness (marked as high, low, none). Variations in roughness were illustrated using 500-ms segments of natural voice stimuli taken from the KEDVD.¹ Participants could hear each stimulus by clicking on the appropriate button. The stimuli were presented binaurally at a comfortable level using headphones (Sennheiser HD 280 Pro or HD 570). After all eight stimuli were played, participants were allowed to replay the stimuli as often as necessary to identify each percept correctly. Following this brief task (roughly 5–10 min), participants completed the MF task as described in the SINE experiment.

Participants performed the matching test immediately after completing both familiarization exercises. These test procedures were identical to the matching task of the SINE experiment with the exception of the comparison stimulus. The stimulus set included the same standards used in the rating and SINE experiments. The test equipment and environment were identical to those described in the SINE experiment.

Results

The mean thresholds obtained with the SQUARE comparison are shown as a function of decreasing roughness in Figure 5 (middle panel). Corresponding boxplots of the thresholds are shown in the middle panel of Figure 6. Listener thresholds varied from –24.6 dB (least roughness) to –10.5 dB (most roughness), resulting in a similar overall range of thresholds as the SINE thresholds (14.1 dB and 13.7 dB, respectively). This indicates that the two comparisons yielded a similar degree of differentiation among the perceived roughness of the current set of reference stimuli. The use of a square-wave modulator, however, resulted in a shift in the mean thresholds to lower values of the modulation index than that obtained with the sine-wave modulator (means of –18.3 dB and –16.2 dB, respectively). This is consistent with the fact that for a given modulation index, square-wave modulation is perceived to be rougher than sine-wave modulation (Zwicker & Fastl, 1990).

After confirming that the listener ratings were normally distributed (a one-sample Kolmogorov-Smirnov test was not significant for any of the voice standards; $p > .05$), reliability for the SQUARE experiment was calculated as the Pearson’s correlation coefficient across (interjudge) and within (intrajudge) participants in a manner identical to that described in the SINE experiment. Average intrajudge reliability was 0.91 ($SD = 0.06$), suggesting a high reliability in judgments within listeners; however, average interjudge reliability was moderate at 0.61 ($SD = 0.06$). The variability in thresholds across all

¹There was no compelling reason to use natural or synthetic stimuli for this demonstration.

standards was examined by regressing the IQRs for each standard onto the listener thresholds (shown in Figure 7; red dashed line and squares). The distribution of IQRs showed an opposite trend across the range of roughness compared with the SINE experiment, as evidenced by the change in direction of the regression slope (-0.245 ; $R^2 = .11$, $p = .051$). The negative slope, although small, indicates greater variability for stimuli matched to a lower modulation index (i.e., less roughness). Although the square-wave modulation resulted in greater roughness of the sawtooth-and-noise carrier, it led to poor discrimination of roughness for stimuli that had very low levels of roughness.

Also, the SQUARE matching thresholds were compared to the rating scale data (shown in the middle panel of Figure 8). The Pearson's correlation between the SQUARE judgments and ratings was high and statistically significant ($r = .876$, $p < .01$). Further, regression analysis revealed a linear relationship between the two tasks ($R^2 = .767$) with a slope of 3.02, indicating that participants were able to judge roughness using the SQUARE comparison stimulus. These data further support the use of a matching task to obtain measures of roughness.

Experiment 4: Roughness Matching Using POWER Modulation

In a final effort to improve reliability and decrease the dispersion across the range of roughness, the comparison was modified once again. The new comparison stimulus presented in this experiment was predicted to serve as a better tool for the comparison of roughness in standards that are both low and high in roughness. The resulting matching thresholds were once again compared to rating scale judgments.

Method

Participants—Ten individuals (seven women, three men; age range = 19–51 years, $M = 22.2$ years) participated in the POWER experiment. One individual had previously participated in the SINE experiment, and two individuals had previously participated in the SQUARE experiment. However, testing for the POWER experiment was conducted at least 2 months after the SINE or SQUARE experiments, so practice effects were likely to be very small. The three listeners with experience had a similar range of thresholds (18.8 dB) to the seven listeners without experience (17.8 dB). Participant inclusion criteria and payment schedule were identical to Experiment 1.

Comparison stimulus—The carrier waveform of the comparison stimulus was identical to that used for the SINE and SQUARE experiments. The modulator was a 25-Hz sine wave raised to the power of four prior to modulation as shown in Equation 4 (see Figure 4 for a sample of the waveform):

$$Y(t) = \{1 + m[\sin(2\pi ft + \theta)]\}^4 * c(t). \quad (4)$$

The result is a temporal envelope that has sharper amplitude peaks and longer amplitude valleys than sinusoidal modulation, and it has a lower frequency that further increased the perception of fluctuation. Based on pilot tests and evidence from Zwicker and Fastl (1990), this modulation function was expected to result in a wider range of roughness for the comparison. It was hypothesized that this modulation function would make it easier for participants to match the roughness of the comparison to the standards.

Procedure—Participants were required to complete three familiarization exercises prior to the main experiment—the “visual sort ranking” (VSR), the IQD, and the MF tasks. The VSR task allowed listeners to practice differentiating among a large set of voice stimuli

varying in roughness. In this task, the 34 standards used in the matching task were displayed as buttons on the right side of a MATLAB GUI. Participants were asked to arrange the buttons vertically in order of roughness (least rough to most rough). Participants could click on a button to hear the standard and move the buttons to form a list on the left side of the GUI. Stimuli were presented diotically at a comfortable level over headphones (Sennheiser HD 280 Pro or HD 570). When the participant was satisfied with the rank order of stimuli, he or she pressed a button marked “Done.” These rankings were then compared against the consensus ranks assigned by a panel of three experts (the second and third authors and a third expert with extensive experience in the assessment and rehabilitation of disordered voices), and feedback was provided to the participants. All stimuli that differed in rank from the expert ranking by five or more units were highlighted in a different color. Two different colors were used—one to indicate that the assigned rank was lower than the expert rank and another that indicated that the assigned ranks were higher than those provided by the experts. Following this feedback, participants were allowed to replay and rearrange the rankings. When finished, the rankings were once again compared to the expert ranks and feedback was provided for the second time. However, the second feedback only indicated when the assigned ranks deviated by more than five from the expert ranks, and no information on the direction of error was provided. Participants were then given one final opportunity to listen and reorder the buttons, before a final submission of their rank order. The final rank order submitted by each participant was compared to the expert rank order. Participants were allowed to proceed to the next task only if the Pearson’s product–moment correlation between their rankings and the expert rankings was greater or equal to 0.70. All participants met this criterion.

This procedure took between 30 min and 2 hr for each participant and was completed only during the first test session for each participant. Afterwards, participants performed the IQD and MF tasks as described in the SINE and SQUARE experiments. Participants began the main test in the same or subsequent test session, depending upon the time remaining in the first session. Although the protocol of this matching test was similar to the SINE and SQUARE experiments in terms of the standard stimuli, environment, and procedures, participants were asked to match the “fluctuation” to the modified comparison stimulus. Also, as previously noted, the modulating signal was a complex power function.

Results

The bottom panel of Figure 5 shows the mean thresholds (with standard error bars) and the bottom panel of Figure 6 shows the boxplots of the thresholds, both in descending order of roughness. Thresholds obtained using the POWER modulator ($M = -25.9$ dB) were shifted even lower than those obtained with the SQUARE modulator ($M = -18.3$ dB). Thresholds ranged from -33.7 dB (least roughness) to -16.8 dB (most roughness). This confirmed that perceived roughness, for a given modulation index, increased from SINE to SQUARE to POWER functions. Matching thresholds estimated with the power function spanned a range of 16.9 dB, greater than either the square-wave (14.1 dB) or the sine-wave (13.8 dB) modulators. This is potentially advantageous, as a greater range of indices offers the possibility of improved differentiation of the roughness among a given set of stimuli.

The use of linear statistics was confirmed using a one-sample Kolmogorov-Smirnov test. Results were not significant for any of the voice standards ($p > .05$). Reliability for the POWER comparison was computed using Pearson’s correlation coefficients within and across individual listeners, as described earlier. The intrajudge reliability was found to be 0.84 ($SD = 0.06$), whereas the interjudge reliability was observed to be 0.68 ($SD = 0.05$). An examination of the dispersion across standards in Figure 7 reveals a negative trend with a slope of -0.274 ($R^2 = .54$, $p = .001$). This indicates that variability decreased with matched modulation depth, similar to the SQUARE experiment.

Pearson's correlation between the POWER thresholds and the ratings suggested that the two judgments were highly correlated ($r = .908$, $p < .01$). These thresholds were best described by a linear relationship, determined through subsequent regression analysis ($R^2 = .83$, $p < .001$) shown in the right panel of Figure 8. A slope of 3.9 was found when matching judgments were regressed on rating estimates, suggesting that the ratings were on a more compressed scale than the POWER thresholds.

Comparison of SINE, SQUARE, and POWER

In the initial analyses described in the experiment sections above, we observed high correlations between roughness estimates obtained from all three modifications of the matching task and rating scale judgments. In the following analysis, the matching thresholds obtained in the three experiments were compared to identify any differences in reliability or the measured perceptual distances between standards. Intrajudge and interjudge reliability were compared using one-way analysis of variance (ANOVA) with Bonferroni corrections. Whereas the main effect of the ANOVA for the intrajudge reliability was borderline significant, $F(2, 22) = 3.478$, $p = .049$, none of the pairwise comparisons were significant. Therefore, the average intrajudge reliability obtained in all three experiments was similarly high. The POWER experiment resulted in the highest average interjudge reliability; however, an ANOVA of the inter-judge reliability was not significant, $F(2, 22) = 2.977$, $p = .072$. Because no clear difference in reliability across experiments was seen, it appears that the familiarization tasks had minimal effect on the average inter- and intrajudge reliability. However, this cannot be confirmed through the present study because two independent variables (familiarization task and modulation function) were co-varied.

The variability in matching task appears to vary as a function of the comparison stimulus (see Figure 7 for a summary). The SINE experiment resulted in greatest variability for stimuli perceived to have the highest degree of roughness. This is likely because the SINE modulation, even at maximum modulation depth, did not result in the same magnitude of roughness as the dysphonic voices. Consequently, listeners were unable to obtain a good match between the two stimuli. The SQUARE and the POWER experiments resulted in comparison stimuli with greater roughness magnitude, thereby lowering the variability for stimuli with high roughness. At low roughness, the high variability may simply reflect larger difference limens for modulation depth discrimination. Similar results were also obtained for breathiness matching by Patel et al. (2012).

Statistical analysis to determine whether different comparison stimuli resulted in different perceptual distances across stimuli was performed by regressing the matching thresholds obtained from each experiment onto one another. Linear regression functions with a slope of unity would indicate that each experiment resulted in equal perceptual distances between stimuli. Regression functions with slopes greater (or less) than 1.00 would indicate that a wider range of matching thresholds was obtained with one comparison for the same stimuli. Figure 9 shows the regression functions comparing the SINE and SQUARE thresholds (left panel), the SINE and POWER thresholds (center panel), and the SQUARE and POWER thresholds (right panel). Note that more negative values represent lower roughness. The slope of the regression line in the left panel (SINE thresholds regressed onto SQUARE thresholds) was near unity (slope = 0.92). In addition, the range of thresholds for each condition was very similar (for SINE, 13.8 dB; for SQUARE, 14.1 dB). This suggests that the two tasks provided a similar resolution of thresholds relatively equal distances across stimuli. On the other hand, the slopes obtained for the comparisons in the center and right panels (SINE-POWER and SQUARE-POWER) were below 1.00 (0.72 and 0.74, respectively), suggesting greater perceptual distances for the POWER condition. This is

further supported by the larger range of thresholds observed for the POWER condition (16.9 dB) compared to the SINE and SQUARE conditions.

An analysis of the mean thresholds also suggested differences among tasks. Statistical results from a one-way ANOVA showed a significant main effect in a test of the difference in mean thresholds across experiments, $F(2, 99) = 50.085$, $p < .01$. Bonferroni-adjusted comparisons indicated that the POWER thresholds ($M = -25.85$ dB) significantly differed from the SINE ($M = -16.18$ dB, $p < .01$) and SQUARE thresholds ($M = -18.27$ dB, $p < .01$). The mean SINE thresholds did not significantly differ from the SQUARE thresholds ($p > .05$). These differences in the absolute magnitude of the matching thresholds likely resulted from modifications to the comparisons. The modulator used in the POWER condition resulted in the perception of more roughness compared to the sine-wave or square-wave modulators at the same modulation depth.

General Discussion

Matching tasks have been commonly used in psychoacoustic experiments to study perceptual attributes of sounds such as loudness and pitch. These tasks are advantageous as they provide ratio-level data and minimize contextual biases resulting from the arbitrary assignment of numbers to quantify the magnitude of a perceived quality. A matching task using a sawtooth-and-noise comparison has also been successfully used by Patel et al. (2012) to obtain matching judgments of breathiness in voice stimuli. This particular comparison was selected because of its spectral and qualitative similarity to natural voices (as compared to a pure tone, for example). The purpose of the experiments presented here was to determine whether the matching task described by Patel et al. (2012) could be modified to quantify the magnitude of roughness in voices and to identify an appropriate comparison (specifically, the modulating function) for this task. A comparison of the matching thresholds to the ratings of the same voice stimuli showed that the judgments were very similar, evidenced by high R^2 values, for all three matching tests. In addition, all three conditions provided moderately high intra- and interjudge reliability. These results were seen despite differences in the modulating function of the comparison stimulus as well as the initial familiarization exercises.

The results of the present study indicate that a matching task using an amplitude-modulated sawtooth-wave-and-noise carrier as the comparison stimulus can be used to obtain judgments of roughness in voice stimuli. These results also validate the use of the matching task in estimating voice quality in natural voices, because previous work has demonstrated the matching task with synthetic vowel stimuli. In general, the reliability of roughness judgments was observed to be lower than that for breathiness judgments reported by Patel et al. (2012). Some variability may be attributed to a smaller number of participants (five participants in the SQUARE condition vs. 10 participants for SINE and POWER); however, the primary source of this discrepancy is not clear. It is possible that perceptual judgments of roughness are inherently more difficult and variable than those for breathiness. Future experiments may nevertheless benefit from including a larger and similar number of participants across tasks, as a low number may affect the strength of regressions and the identification of outliers. Although the addition of multiple familiarization exercises did not seem to affect the reliability of judgments, participant reports suggested that these tasks decreased the difficulty of this task. All three test procedures (SINE, SQUARE, and POWER) resulted in similar average matching thresholds, but the wider range of thresholds obtained in the POWER condition is potentially advantageous for better discrimination in roughness across stimuli. For these reasons, it is recommended that roughness matching be conducted with sinusoidal amplitude modulation raised to a power of four and applied to a sawtooth-and-noise carrier. Some familiarization tasks, such as the MF and IQD tasks, are

recommended when conducting the matching tasks; however, the more time-intensive VSR had little effect on reliability.

Although a matching task for evaluating voice quality has previously been described by Kreiman, Gerratt, and colleagues (Gerratt & Kreiman, 2001; Kreiman et al., 2007), the matching task described here is different in three regards. First, Kreiman, Gerratt, et al.'s approach used a nonspeech stimulus instead of a synthesized vowel as the comparison stimulus. A speech-like stimulus is advantageous because it is easy to generate and manipulate while it closely approximates natural vowels, and it allows the experimenter to focus listener attention on a single aspect or dimension of interest. The simplicity of the stimulus minimizes concerns about confounding factors in natural or synthetic vowels, such as vowel formant frequencies or formant bandwidths. Second, unlike Gerratt and Kreiman (2001), the present approach permits manipulation of a single acoustic variable. Although an approach that allows adjustment of several variables at once may or may not result in a better qualitative match, such a method also permits multiple parameter combinations to reflect identical percepts. There is no easy way to equate these settings across listeners or experiments, and as a result, it is difficult to compare and model results. In contrast, allowing participants to manipulate only one parameter enables an experimenter to quantify changes in matching thresholds easily. This allows easy comparison of data across experiments, listeners, laboratories, or other experimental variables. Such data also lends itself for the development and validation of computational models of voice quality. Finally, the present research seeks to obtain a stable average measurement of voice quality rather than achieve high agreement across participants. As with many physical and psychological measurements, it is assumed that some variability is inherent to behavior and to the process of measurement. By virtue of being more stable than individual data, the average measure also has greater relevance, such as when used as a measure of treatment outcome. Therefore, this technique seeks to obtain the central tendency of the distribution of matching thresholds.

For the purposes of this research, we have adopted fairly rigorous psychophysical methods to provide robust data on which to base theory and models. The experiments described here required several hours of listening from each participant. The time-intensive nature of this task makes it difficult to use it in everyday clinical practice. However, the data obtained through this psychophysical scaling approach have properties that make them suitable for use in model development. The eventual goal of this work is to generate a model for the perception of dysphonic voice quality, which can then be used to develop clinically useful and feasible approaches for the measurement of voice quality. Nevertheless, in specific cases where test time and availability of listeners are not limiting factors, the single-variable matching task could be used to obtain accurate and unbiased measures of voice quality for clinical use as well.

Conclusions

A simple matching task to quantify roughness in voice stimuli is described. This task allows measurement of roughness using a fixed measurement unit (modulation depth in dB or percent). Three adaptations of the matching task proposed by Patel et al. (2012) were implemented, each differing in the amount of initial familiarization tasks given to participants and the properties of the comparison stimulus. A comparison of the thresholds with rating scale judgments obtained in a preliminary experiment suggested that participants in all three matching tests were able to evaluate the roughness of natural voice stimuli successfully. Therefore, any of these three matching conditions could be used to obtain judgments of roughness. Because an increased range of thresholds and a relatively balanced distribution of variance were seen for the matching thresholds that used an exponential amplitude modulation function, we recommend using the POWER comparison. The nature

and duration of the initial familiarization exercises used in the present experiments did not appear to have a significant effect on the reliability of the matching task. As such, the additional VSR task does not prove beneficial, given the additional 30 min required to perform this task. Hence, future work may benefit by using the POWER comparison with matching training (i.e., the MF task) and a brief quality demonstration (similar to the IQD exercise) prior to obtaining judgments of roughness.

Acknowledgments

This research was supported by Grant NIH R01 DC009029 from the National Institute on Deafness and Other Communication Disorders. We would like to thank Stacie Cummings for her help with data collection and Judith Wingate for lending her ears.

References

- American National Standards Institute. ANSI S3.21-2004. New York, NY: Author; 2004. Methods for manual pure-tone threshold audiometry.
- American National Standards Institute. ANSI S3.4-2007. New York, NY: Author; 2007. Procedure for the computation of loudness of steady sounds.
- Belkin K, Martin R, Kemp SE, Gilbert AN. Auditory pitch as a perceptual analogue to odor quality. *Psychological Science*. 1997; 8:340–342.
- de Krom G. Some spectral correlates of pathological breathy and rough voice quality for different types of vowel fragments. *Journal of Speech and Hearing Research*. 1995; 38:794–811. [PubMed: 7474973]
- Deal RE, Emanuel FW. Some waveform and spectral features of vowel roughness. *Journal of Speech, Language, and Hearing Research*. 1978; 21:250–264.
- Gerratt BR, Kreiman J. Measuring vocal quality with speech synthesis. *The Journal of the Acoustical Society of America*. 2001; 110:2560–2566. [PubMed: 11757945]
- Gescheider, GA. *Psychophysics: Method and theory*. Hillsdale, NJ: Erlbaum; 1976.
- Guilford, JP. *Psychometric methods*. New York, NY: McGraw-Hill; 1954.
- Hirano, M. *Clinical examination of voice*. New York, NY: Springer; 1981.
- Kempster GB, Gerratt BR, Verdolini-Abbott K, Barkmeier-Kraemer J, Hillman RE. Consensus auditory-perceptual evaluation of voice: Development of a standardized clinical protocol. *American Journal of Speech-Language Pathology*. 2009; 18:124–132. [PubMed: 18930908]
- Kewley-Port D, Pisoni DB, Studdert-Kennedy M. Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants. *The Journal of the Acoustical Society of America*. 1983; 73:1779–1793. [PubMed: 6223060]
- Kreiman J, Gerratt B, Berke GS. The multidimensional nature of pathological voice quality. *The Journal of the Acoustical Society of America*. 1994; 96:1291–1302. [PubMed: 7962996]
- Kreiman J, Gerratt B, Ito M. When and why listeners disagree in voice quality assessment tasks. *The Journal of the Acoustical Society of America*. 2007; 122:2354–2364. [PubMed: 17902870]
- Martin D, Fitch J, Wolfe V. Pathologic voice type and the acoustic prediction of severity. *Journal of Speech and Hearing Research*. 1995; 38:765–771. [PubMed: 7474970]
- Martin D, Wolfe V. Effects of perceptual training based upon synthesized voice signals. *Perceptual and Motor Skills*. 1996; 83:1291–1298. [PubMed: 9017742]
- Parducci A, Wedell DH. The category effect with rating scales: Number of categories, number of stimuli, and method of presentation. *Journal of Experimental Psychology: Human Perception and Performance*. 1986; 12:496–516. [PubMed: 2946806]
- Patel S, Shrivastav R, Eddins DA. Perceptual distances of breathy voice quality: A comparison of psychophysical methods. *Journal of Voice*. 2010; 24:168–177. [PubMed: 19185451]
- Patel S, Shrivastav R, Eddins DA. Developing a single comparison signal for matching breathy voice quality. *Journal of Speech, Language, and Hearing Research*. 2012; 55:639–647.
- Poulton, EC. *Bias in quantifying judgments*. Hove, United Kingdom: Erlbaum; 1989.

- Shrivastav, R. Evaluating voice quality. In: Ma, E.; Yu, E., editors. Handbook of voice assessments. San Diego, CA: Plural; 2011. p. 305-318.
- Shrivastav R, Camacho A, Patel S, Eddins DA. A model for the prediction of breathiness in vowels. The Journal of the Acoustical Society of America. 2011; 129:1605–1615. [PubMed: 21428523]
- Shrivastav R, Sapienza CM, Nandur V. Application of psychometric theory to the measurement of voice quality using rating scales. The Journal of Speech and Hearing Research. 2005; 48:323–335.
- Siegel W. Memory effects in the method of absolute judgment. Journal of Experimental Psychology. 1972; 94:121–131.
- Stevens SS. On the theory of scales of measurement. Science. 1946 Jun 7.103:677–680.
- Stevens, SS. Psychophysics: Introduction to its perceptual, neural, and social prospects. New York, NY: Wiley; 1975.
- Stevens JC, Hall JW. Brightness and loudness as functions of stimulus duration. Perception & Psychophysics. 1966; 1:319–327.
- Stevens SS, Volkman J. The relation of pitch to frequency: A revised scale. The American Journal of Psychology. 1940; 53:329–353.
- Stewart N, Brown GDA, Chater N. Absolute identification by relative judgment. Psychological Review. 2005; 112:881–911. [PubMed: 16262472]
- Thurstone LL. A law of comparative judgment. Psychological Review. 1927; 34:273–286.
- Toner MA, Emanuel FW. Direct magnitude estimation and equal appearing interval scaling of vowel roughness. Journal of Speech and Hearing Research. 1989; 32:78–82. [PubMed: 2704204]
- Zwicker, E.; Fastl, H. Psychoacoustics: Facts and models. Berlin, Germany: Springer-Verlag; 1990.

Figure 1. Ratings of the 34 voice standards. Bars represent standard error of the mean. Note that the x -axis represents the numerical label assigned to each standard stimulus, where 1 = *minimal roughness* and 5 = *severe roughness*.

Figure 2.

Boxplots of the 34 voice standards. Outlier symbols are filled circles and stars (in addition to the listener number), which represented 1.5 and 3 times the interquartile range (IQR), respectively.

Figure 3. Distribution of the interquartile range across listener ratings for all standard stimuli. The solid line represents the regression line (RL).

Figure 4.

Waveform segments illustrating the amplitude modulation waveforms used in creating the comparison stimuli for the three roughness matching experiments (a 40-Hz sine wave, a 40-Hz square wave, and a 25-Hz sine wave raised to a power of four). Each of these modulators was applied to a sawtooth-and-noise complex as described in the text.

Figure 5. Matching thresholds for 34 voice stimuli of the SINE experiment (top panel), SQUARE experiment (middle panel), and POWER experiment (bottom panel). Bars represent standard error of the mean. Note that the x -axis represents the stimulus identification number for the standards. The 34 standards are arranged in descending order of perceived roughness for each task. To compare threshold values across tasks (and the ratings task), the standards are labeled in each graph according to the ratings task.

Figure 6.

Boxplots of the 34 voice standards of the SINE experiment (top panel), SQUARE experiment (middle panel), and POWER experiment (bottom panel). Outlier symbols are filled circles and stars (in addition to the listener number), which represented 1.5 and 3 times the IQR, respectively. Note that the x -axis represents the stimulus identification number for the standards. The 34 standards are arranged in descending order of perceived roughness for each task. To compare threshold values across tasks (and the rating task), the standards are labeled in each graph according to the rating task.

Figure 7.

Distribution of the interquartile ranges (IQRs) for matching thresholds (in dB) in the SINE, SQUARE, and POWER tasks. The blue solid, red dotted, and black dashed lines represent the linear regression (LR) line for the SINE, SQUARE, and POWER conditions, respectively.

Figure 8.

Matching thresholds (in dB) of the 34 standards regressed onto the ratings for the SINE experiment (left panel), SQUARE experiment (center panel), and POWER experiment (right panel). The solid line is the linear regression line.

Figure 9.

Scatterplot of thresholds (in dB) obtained for each pair of matching tests for 34 standards: SQUARE thresholds regressed onto the SINE thresholds (left panel), SINE thresholds regressed onto the POWER thresholds (center panel), and SQUARE thresholds regressed onto the POWER thresholds (right panel). The solid line is the linear regression line.