Psychometric properties associated with perceived vocal roughness using a matching task

David A. Eddins^{a)}

Department of Communication Sciences and Disorders, Global Center for Hearing and Speech Research, University of South Florida, 4202 East Fowler Avenue, PCD 1017, Tampa, Florida 33620
deddins@usf.edu

Rahul Shrivastav^{b)}

Malcolm Randall VAMC, 1601 S.W. Archer Road, Gainesville, Florida, 32608-1197
rahul@msu.edu

Abstract: A psychophysical matching paradigm has been used to better quantify voice quality under laboratory conditions. The goals of this study were to establish which of two candidate comparison stimuli would best ensure that the range of perceived vocal roughness could be adequately bracketed using a matching task and to provide a general solution to the problem of estimating vocal roughness. Psychometric functions for roughness matching indicated that a speech-like sawtooth-plus-noise complex (20 dB signal-to-noise ratio) amplitude modulated by a sinusoidal function raised to the 4th power yielded a comparison stimulus with a perceptual dynamic range well suited for roughness matching.

© 2013 Acoustical Society of America

PACS numbers: 43.71.Bp, 43.71.Gv, 43.72.Ar [AC]

Date Received: May 2, 2013 **Date Accepted:** August 9, 2013

1. Introduction

Information about talker identity, state, emotion, and health may be carried by voice quality. Though more than 65 attributes of voice quality have been described,¹ three primary dimensions are breathiness, roughness, and strain.² Despite broad agreement on the importance of these dimensions, the psychophysical methods used to estimate perceived voice quality continue to pose barriers in both clinical practice and basic research. The goal of the current study is to establish the suitability of a psychophysical matching task that may be broadly adopted as a standard method of estimating perceived roughness of voiced stimuli.

Similar to other voice quality dimensions, previous investigations of roughness have relied largely on rating-scale³ and magnitude estimation⁴ tasks. These psychophysical methods are relatively easy to implement; however, the data they provide are difficult to interpret for a number of reasons that have been detailed extensively in the psychophysics literature.⁵ Specifically, the arbitrary assignment of numbers to denote the magnitude of a percept in the rating scale or magnitude estimation tasks results in contextual biases and poor reliability of perceptual judgments. By avoiding these problems, a matching task is better suited for the perceptual study of voice quality.^{6,7} Furthermore, perceptual judgments from a single-variable matching task provide ratio-level data, as opposed to the ordinal data from rating scales, permitting more detailed mathematical treatment of the data.⁸

^{a)} Author to whom correspondence should be addressed.

^{b)} Also at: Department of Communicative Sciences and Disorders, Michigan State University, 1026 Red Cedar Road, East Lansing, Michigan 48824.

Two kinds of matching tasks have been proposed to study voice quality. Gerratt and Kreiman⁹ and Kreiman *et al.*¹⁰ proposed a multi-parameter matching task to quantify voice quality. In this approach, a listener manipulates multiple parameters of a speech-synthesizer to replicate a target voice stimulus. The set of synthesizer parameters that provides a perceptual match to the target voice stimulus is used to index voice quality. In contrast, Patel *et al.* have described a single-variable adaptive matching paradigm to quantify voice quality under laboratory conditions. This approach has been tested for breathy^{6,11} and rough⁸ qualities. In this task, the magnitude of the voice quality of interest (e.g., roughness) is measured by comparing the test stimulus (the “standard”) to a single reference stimulus (the “comparison”). Standard stimuli consist of natural or synthetic voice samples that vary in voice quality and the comparison stimulus is an artificial, non-speech stimulus that can be varied along the voice quality dimension of interest. The independent variable is a single parameter of the comparison stimulus and that parameter is manipulated in the context of the matching task until the quality of the comparison matches that of the standard. That value represents the point of subjective equality between the standard and comparison stimuli and is taken as an index of the voice quality percept under study. The value of the independent variable is manipulated following each listener response using a method of limits, resulting in a series of ascending and descending trials that lead to a single estimate of the vocal quality percept. The use of a well-defined comparison stimulus makes it possible to compare data across multiple experiments, participants, and laboratories. Although relatively novel to the study of dysphonic voice quality, this approach is routinely used in nearly all aspects of sensation and perception.^{12,13}

In an initial effort to provide a general solution to the measurement of perceived vocal roughness, Patel *et al.*⁸ explored three different candidate comparison stimuli and concluded that an amplitude-modulated sawtooth wave provided reliable estimates of perceived vocal roughness. For the adaptive matching task described above to work well, however, the independent variable must be capable of producing a range of perceived sound quality that exceeds the range of sound quality inherent in the standard stimulus set under study. In other words, adjustment of the independent variable toward one end of the continuum should lead to stimuli that are perceived to be more rough than the standard and adjustment toward the other end should lead to stimuli perceived to be less rough than the standard. When this is true, the point of subjective equality can be bracketed and estimated. In developing the comparison stimulus, Patel *et al.*⁸ noted that the use of a sinusoidal modulator produced insufficient roughness in the comparison stimulus to adequately match the perceived roughness of the most rough standard stimuli. The perceived roughness of the comparison stimulus was increased in successive experiments by using a sine function raised to the power of either two or four. This has the effect of sharpening the peaks and broadening the valleys of the modulation function, increasing the fourth moment of the stimulus envelope¹⁴ from 1.75 for the sine function to 2.37 for the power of two and 2.61 for the power of four functions. In practice, the two power functions appeared to provide adequate roughness for the task but there was no clear basis for deciding between the two. The perceived smoothness (opposite of roughness) of the comparison stimulus was increased in successive experiments by increasing the signal (modulated sawtooth wave) to noise ratio from 12 to 20 dB.

In an effort to provide a standard method for estimating perceived roughness, it would be advantageous to be able to distinguish between the two candidate comparison stimuli and to be able to recommend a single comparison stimulus that meets the primary requirements of the matching task. To do so, the present study uses a method of constant stimulus rather than an adaptive procedure for obtaining matching thresholds. This will allow one to determine the psychometric function relating perceived roughness to modulation depth for each standard stimulus in a set. If the resulting psychometric functions extend well below and above the point of subjective equality as the modulation depth is varied, then one can conclude that the comparison stimulus

has an adequate range of perceived roughness for the matching task. If not, another comparison stimulus is required. Here we test the hypothesis that the increased envelope variability of the sinusoidal modulator raised to a power of four combined with the increased signal to noise ratio of that comparison envelopes a broader range along the roughness continuum, thereby providing a more suitable comparison stimulus for a set of stimuli that varies from extremely smooth to extremely rough.

2. Method

2.1 Listeners

Ten listeners ranging in from 18 to 25 yr (mean 22 yr) with normal hearing¹⁵ (<20 dB hearing level (HL) from 250 to 8000 Hz) participated. All listeners were native speakers of American English. Listeners were paid an hourly wage for their participation. All procedures were approved by appropriate Institutional Review Board and all listeners consented to participate.

2.2 Stimuli

2.2.1 Standard stimuli

In the first set of measurements, 34 voices were selected from the Kay Elemetrics Disordered Voice Database (KEDVD; Kay Elemetrics, Inc., Lincoln Park, NJ) using stratified sampling 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. Stimuli consisted of a 500-ms segment excised from the temporal center of the original waveform. To reduce onset and offset artifacts during playback, a 10-ms cosine window was used to shape the stimulus onset and offset. These stimuli are referred to as “standard” stimuli, borrowing from common nomenclature in used in psychoacoustic experiments. For the second set of measurement, the set of 34 was reduced to 10 voices selected to represent a wide range of perceived roughness and to span the range of psychometric function parameters obtained with the full set of 34 voices.

2.2.2 Comparison stimuli

The two comparison stimuli used here are based on the comparison stimuli used by Patel *et al.*^{6,8} The carrier was a sawtooth wave with a fundamental frequency of 151 Hz that was low-pass filtered and mixed with noise. In one condition, the carrier was amplitude modulated with a sine function raised to the power of two, hereafter termed the “SQUARE” condition. In a second set of experiments, the carrier was amplitude modulated with a sine function raised to the power of four, here after termed the “QUAD” condition. For the SQUARE condition, the fundamental frequency (151 Hz) and low-pass filter characteristics (slope of -7 dB per octave, cutoff of 151 Hz) were chosen based on the average parameters computed from the entire set of disordered voices in the KEDVD. Noise with the same spectral shape as the filtered sawtooth was added at an arbitrarily defined 12-dB signal-to-noise ratio (SNR) to make the comparison sound more speech-like. For the QUAD condition, a low-pass filter slope of -12 dB/octave was chosen to maximize perceived roughness and an SNR of 20 dB was chosen to maintain a noise-like quality similar to the SQUARE comparison. The waveform of the sawtooth-plus-noise complex was amplitude modulated as shown in Eq. (1):

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

where m is modulation depth (0 to 1), f is modulation frequency in Hz (fixed at 25 Hz), t is time in seconds, θ is starting phase (fixed at 0 radians), p is power (either 2 or 4), 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.

2.3 Procedure

A two interval, two-alternative, forced-choice procedure was used. The standard (voice stimulus) being evaluated was presented on the first interval of the trial and the comparison (modulated carrier) stimulus was presented on the second interval of the trial. The participant indicated via button press which of the two was perceived to have greater roughness. The amplitude modulation depth (i.e., the independent variable) of the comparison ranged from -11 dB (most rough) to -39 dB (least rough) in steps of 2 dB. Because of an error in data collection, matching thresholds for modulation depths of -17 dB and -27 dB were not collected. This left 13 values along the x axis, which is more than sufficient to estimate a robust psychometric function given the 2-dB step size chosen. Each pair of sounds (standard speech token, comparison modulation depth) was presented 10 times in random order for a total of $34 \text{ voices} \times 13 \text{ depths} \times 10 \text{ repetitions} = 4420$ trials. In the second measurement set, the number of standard voices was reduced to 10 and the number of modulation depths was reduced to 11 (-11 to -33 dB in 2-dB steps with -27 dB erroneously omitted during data collection) resulting in a total of $10 \times 11 \times 10 = 1100$ trials for the second and third measurement sets. The percentage of “greater” roughness judgments was computed for each standard voice, depth, and listener allowing a psychometric function to be generated for each standard voice and each listener. Testing required 5 to 8 h over 3 to 5 sessions lasting 1.5 to 2 h each.

2.4 Equipment

Stimulus generation, presentation, and response collection were controlled by TDT (Tucker-Davis Technologies) SykofizX software and TDT modules including an enhanced real-time processor (TDT RP2.1), programmable attenuators (TDT PA5), and headphone buffer (TDT HB7). All stimuli were re-sampled at 24 414 Hz to accommodate the allowable sample rates of the TDT hardware. Stimuli were presented monaurally (right ear, chosen arbitrarily) at a level of 85 dB sound pressure level (SPL) via Etymotic ER-2 insert earphone calibrated using a G.R.A.S. IEC 126 2-cc coupler connected to a Larson-Davis (model 2800) sound level meter (linear weighting). All testing was conducted in an Industrial Acoustics Corporation sound attenuating chamber.

2.5 Psychometric functions

Psychometric functions for individual talkers and individual listeners resembled a classic sigmoidal shape. Raw functions for each subject and each condition were fit with a model representing the logistic distribution as illustrated in Eq. (2):

$$p(c) = 1/[1 + e^{-(x-x_0)/b}], \quad (2)$$

where $p(c)$ is the probability of the standard being judged rougher than the comparison, b is the growth constant (i.e., steepness factor) of the function, X_0 is the midpoint (50% correct) of the function, and x is the modulation depth value. Each psychometric function was fit using the unconstrained nonlinear optimization method implemented in the MATLAB `fminsearch` function. The algorithm searched for parameters that minimized the sum of the squared differences between the behavioral and predicted data using an iterative process.

3. Results

Judgments were similar across observers and are well represented by the mean data as shown by the 34 functions in Fig. 1(A). These data are for the SQUARE comparison. The percent of judgments for which the standard was judged rougher than the comparison is displayed on the y axis and modulation depth on the x axis. All functions have

Fig. 1. (Color online) Psychometric functions for roughness perception. The percentage of trials the standard stimulus was judged to be rougher than the comparison stimulus is plotted on the y axis as a function of amplitude modulation depth (in dB) on the x axis. Each datum point represents the mean across 10 listeners. Stimuli represented by psychometric functions shifted to the right (smaller modulation depths) are perceived as less rough than stimuli represented by functions shifted to the left. (A) Functions were obtained using the SQUARE comparison stimulus for each of 34 voice tokens (individual functions). (B) Data for 10 of the 34 voice tokens in panel A are re-plotted from to facilitate comparison to the same 10 voice tokens used to evaluate the QUAD comparison shown in panel C. (C) Functions for the same 10 voice tokens in Panel B evaluated using the QUAD comparison stimulus. Panels A and C display statistics associated with the fitting parameters obtained using Eq. (2).

a characteristic ogive shape but differ in their position along the x axis, slope, and asymptotic performance level. The modulation depth corresponding to the 50% point (midpoint) and the slope were estimated for each of the 34 voices and 10 subjects (340 functions) from the fitted functions according to Eq. (2). Of the 340 psychometric functions, 8 resulted in poor fits (due to nonmonotonicities in the raw data) and were excluded from the summary data below. As summarized in the table-inset to Fig. 1, panel A, midpoints for the SQUARE comparison ranged from -22.3 to -29.6 across voices. The slopes were estimated from the fitted functions according to Eq. (2) and ranged from 1.1 to 2.6 across voice tokens. Note that asymptotic performance beyond modulation depths of -29 dB reflects the smooth end of the roughness continuum for the comparison stimulus. Modulation detection threshold for the comparison stimulus is about -30 dB, indicating that beyond -30 dB, no modulation can be detected.

The psychometric functions for several stimuli asymptote well below the 100% point. In these cases, the standard voice token approaches the minimum roughness of the comparison stimulus. In other words, for the least rough voices, the comparison carrier stimulus is judged as more rough even with minimal superimposed amplitude modulation. Ideally, the comparison stimulus could be judged as rougher than the roughest standard and smoother than the smoothest standard by just varying the depth of amplitude modulation. To potentially achieve this goal, the QUAD comparison stimulus was evaluated. Changes in the modulation function, spectral slope, and SNR were made and informal listening indicated that the QUAD modulation function, -12 dB/octave filter slope, and 20 dB SNR resulted in a stimulus that was rougher than the SQUARE stimulus for large modulation depths and smoother than the SQUARE stimulus for small modulation depths.

To evaluate the suitability of the QUAD comparison stimulus in a formal manner, a direct comparison of the SQUARE and QUAD comparison stimuli was undertaken. To reduce test time, the set of standard voices was reduced from 34 to 10 and the range of modulation depths evaluated was reduced by eliminating the three smallest modulation depths (-32 , -35 , and -38 dB) as described above. Figure 1(B) displays the resulting psychometric functions for the SQUARE comparison. These data are taken directly from Fig. 1(A). Visual comparison of Figs. 1(A) and 1(B) show that the smaller set of 10 voices in Fig. 1(B) result in a family of psychometric functions that are representative of the larger set of 34 voices in Fig. 1(A). Figure 1(C)

displays the psychometric functions obtained for the QUAD comparison stimulus. Based on the fitted functions, midpoints ranged from -19.0 to -26.8 dB across voices while the slopes ranged from 0.7 to 3.6 across voice tokens, as summarized in the table-insert in Fig. 1(C). Of the 100 functions, 4 resulted in poor fits and were excluded from this summary. Midpoints for the QUAD comparison [Fig. 1(C)] are shifted toward greater modulation depths by about 3 to 4 dB relative to the SQUARE comparison, indicating that the QUAD comparison is judged as smoother than the SQUARE comparison. Furthermore, psychometric functions for all standards showed an asymptote at a point above 75%, indicating that the QUAD comparison indeed was smooth enough to be judged as less rough than any of the standard stimuli yet with enough modulation depth could be judged as more rough than any of the standard stimuli.

4. Discussion and conclusions

Previous reports^{6,8,9,11} have detailed several advantages of using a psychophysical matching task for measuring perceived voice quality across a range of stimuli and quality dimensions (i.e., breathiness, roughness). A factor critical to the success of the adaptive matching task is the range of roughness that can be supported by the comparison stimulus. By virtue of manipulating the independent variable (e.g., amplitude modulation depth), the comparison should be perceived as substantially less rough or more rough than any standard voice stimulus one might choose. When this criterion is met, the matching task can include both ascending and descending matching threshold runs by setting the initial independent variable value above or below the value of the final matching threshold. Combining ascending and descending runs helps to avoid any potential directional biases in the matching task. This was evaluated by selecting from the KEDVD a set of voices characterized by a very wide range of perceived roughness and then measuring the psychometric functions resulting from direct comparison of the roughness of the standard and the comparison stimuli. The results from the first set of measurements indicated that the SQUARE comparison stimulus was adequate at the rough end of the roughness continuum but was insufficient at the less rough end of that continuum. On the contrary, psychometric functions for the same standard stimuli and same listeners revealed that the QUAD comparison stimulus could be manipulated to be suitably rough and suitably smooth to serve as a matching stimulus for standards considered to be extremely rough as well as those considered to have no roughness at all. Thus, the hypothesis put forward in the introduction is confirmed by the present experiments, establishing that the psychometric properties of the QUAD comparison stimulus, based on amplitude modulation of a carrier stimulus modeled after a large set of voices from the KEVED, are robust and suitable for use in a psychophysical matching task designed to quantify the perceived roughness of sustained /a/ vowels produced spanning the range from minimal to extreme perceived roughness. The single-variable matching task illustrated here, combined with the QUAD stimulus described in the Methods, provides a general solution to quantifying rough voice quality that can be implemented in any laboratory and that should provide behavioral data that would be easily comparable to the data presented here or any other data obtained with the same comparison stimulus.

Acknowledgment

Work supported in part by a grant from NIH (Grant No. R01 DC009029).

References and links

- ¹M. Pannbacker, "Classification systems of voice disorders: A review of the literature," *Language Speech Hearing Serv. Schools* **15**, 169–174 (1984).
- ²M. Hirano, "Psycho-acoustic evaluation of voice," in *Clinical Examination of Voice* (Springer-Verlag, New York, 1981), p. 81.
- ³G. de Krom, "Some spectral correlates of pathological breathy and rough voice quality for different types of vowel fragments," *J. Speech Hear. Res.* **38**, 794–811 (1995).

- ⁴M. A. Toner, F. W. Emanuel, and D. Parker, "Relationship of spectral noise levels to psychophysical scaling of vowel roughness," *J. Speech Hear. Res.* **33**(2), 238–244 (1990).
- ⁵N. A. MacMillan and C. D. Creelman, *Detection Theory: A User's Guide* (Psychology Press, New York, 2005).
- ⁶S. A. Patel, R. Shrivastav, and D. A. Eddins, "Perceptual distances of breathy voice quality: A comparison of psychophysical methods," *J. Voice* **24**, 168–177 (2011).
- ⁷J. Kreiman, B. Gerratt, and M. Ito, "When and why listeners disagree in voice quality assessment tasks," *J. Acoust. Soc. Am.* **122**, 2354–2364 (2007).
- ⁸S. A. Patel, R. Shrivastav, and D. A. Eddins, "Identifying a comparison for matching roughness," *J. Speech Lang. Hear. Res.* **55**, 1407–1422 (2012).
- ⁹B. R. Gerratt and J. Kreiman, "Measuring voice quality with speech synthesis," *J. Acoust. Soc. Am.* **110**, 2560–2566 (2001).
- ¹⁰J. Kreiman, N. Antónanzas-Barroso, and B. R. Gerratt, "Integrated software for analysis and synthesis of voice quality," *Behav. Res. Methods.* **42**(4), 1030–1041 (2010).
- ¹¹S. A. Patel, R. Shrivastav, and D. A. Eddins, "Developing a single reference signal for matching breathy voice quality," *J. Speech Lang. Hear. Res.* **55**, 639–647 (2012).
- ¹²K. Belkin, R. Martin, S. E. Kemp, and A. N. Gilbert, "Auditory pitch as a perceptual analogue to odor quality," *Psychol. Sci.* **8**, 340–342 (1997).
- ¹³J. C. Stevens and J. W. Hall, "Brightness and loudness as functions of stimulus duration," *Percept. Psychophys.* **1**, 319–327 (1966).
- ¹⁴W. M. Hartmann and J. Pumplin, "Periodic signals with minimal power fluctuations," *J. Acoust. Soc. Am.* **90**, 1986–1999 (1991).
- ¹⁵ANSI S3.21-2010: *Methods for manual pure-tone threshold audiometry* (American National Standards Institute, New York, 2010).