The psychophysics of roughness applied to dysphonic voice

David A. Eddins^{a)}

Department of Communication Sciences and Disorders, University of South Florida,
4202 East Fowler Avenue, PCD 1017, Tampa, Florida 33620, USA

Lisa M. Kopf

Department of Communicative Sciences and Disorders, Michigan State University, 1026 Red Cedar Road,
Room 109, Oyer Speech and Hearing Building, East Lansing Michigan 48824, USA

Rahul Shrivastav

Office of the Vice President for Instruction, University of Georgia, 0309 New College, Athens, Georgia 30602,
USA

(Received 1 August 2015; revised 19 November 2015; accepted 25 November 2015; published online 23 December 2015)

Roughness is a sound quality that has been related to the amplitude modulation characteristics of the acoustic stimulus. Roughness also is considered one of the primary elements of voice quality associated with natural variations across normal voices and is a salient feature of many dysphonic voices. It is known that the roughness of tonal stimuli is dependent on the frequency and depth of amplitude modulation and on the carrier frequency. Here, it is determined if similar dependencies exist for voiced speech stimuli. Knowledge of such dependencies can lead to a better understanding of the acoustic characteristics of vocal roughness along the continuum of normal to dysphonic and may facilitate computational estimates of vocal roughness. Synthetic vowel stimuli were modeled after talkers selected from the Satloff/Heman-Ackah disordered voice database. To parametrically control amplitude modulation frequency and depth, synthesized stimuli had minimal amplitude fluctuations, and amplitude modulation was superimposed with the desired frequency and depth. Perceptual roughness judgments depended on amplitude modulation frequency and depth in a manner that closely matched data from tonal carriers. The dependence of perceived roughness on amplitude modulation frequency and depth closely matched the roughness of sinusoidal carriers as reported by Fastl and Zwicker [(2007) *Psychoacoustics: Facts and Models*, 3rd ed. (Springer, New York)]. © 2015 Acoustical Society of America. [<http://dx.doi.org/10.1121/1.4937753>]

[ICB]

Pages: 3820–3825

I. INTRODUCTION

The concept of vocal roughness is familiar to most people and commonly occurs when someone with an otherwise normal voice has an upper respiratory infection, is fatigued, or is just waking from a restful night's sleep. In clinical cases, an abnormally rough voice may be a consequence of vocal pathology or a functional voice disorder. Most perceptual evaluations of dysphonia include estimates of perceived vocal roughness using clinical instruments like the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V; [Kempster et al., 2009](#)) and the Grade, Roughness, Breathiness, Asthenia, Strain assessment (GRBAS; [Hirano, 1981](#)). Scales designed for clinical voice evaluation, such as the CAPE-V and GRBAS, evaluate roughness using psychophysical rating methods (e.g., Likert or visual analog scales) for the purpose of characterizing voice quality during voice assessment and before, during, or after treatment. In laboratory investigations of vocal roughness, single-variable matching tasks have been

used to determine the relation between perception and physical units associated with the matching stimulus.

The perception of roughness is not specific to voiced speech sounds. Indeed, it is a pan-sensory perceptual phenomenon. Many sounds can be gauged on scales from smooth to rough or less rough to more rough. Beyond sound, roughness is a fundamental concept in mechanoreception (the sense of touch) and much of what is known about roughness has been explored in that context. In the context of sound, in general, roughness has been associated with relatively slow fluctuations (e.g., 15–100 Hz) in the temporal envelope of a stimulus ([Terhardt, 1974](#); [Fastl and Zwicker, 2007](#)¹). In the context of dysphonic voices, roughness has been defined as the psychoacoustic impression of irregularity of vocal fold vibrations ([Hirano, 1981](#)).

Quantification of the perceived roughness of sound was championed by [Fastl and Zwicker \(2007](#); see also [Zwicker and Fastl, 1990](#) and [Fastl and Zwicker, 1999](#)) as detailed in their landmark work on auditory perception. They quantified the perceived roughness of arbitrary sounds using an anchored magnitude estimation task in which the roughness of the standard sound (sound of interest) was compared to the roughness of a 1000 Hz tone that was amplitude modulated by a sine wave with a modulation frequency of 70 Hz and a modulation depth of 100%

^{a)}Also at: Department of Communicative Sciences and Disorders, Michigan State University, 1026 Red Cedar Road, Room 109, Oyer Speech and Hearing Building, East Lansing, MI 48824. Electronic mail: deddins@usf.edu

($m = 1$, where m is the modulation index from 0 to 1) and presented at 60 dB sound pressure level (SPL). The term given to the unit of roughness applied to sound was *asper* meaning “rough” in Latin.

Through a series of perceptual experiments, Fastl and Zwicker (2007) estimated the dependence of perceived roughness on fluctuation rate using tonal carriers that were amplitude modulated at various rates of modulation. A modulation index of 1.0 produces maximum modulation depth and an index of 0.0 produces minimal (no) modulation depth. In the work of Fastl and Zwicker (2007), the depth of modulation was fixed at a maximal depth of 1.0 and perceived roughness magnitude was estimated as a function of modulation rate. Those data are re-plotted in Fig. 1 for 125 Hz, 1000 Hz, and 8000 Hz tonal carriers with the magnitude of perceived roughness expressed in units of aspers on the ordinate and modulation frequency on the abscissa. Perceived roughness initially increases with increasing modulation frequency and reaches a peak of maximal perceived roughness, beyond which perceived roughness declines with further increases in modulation frequency. This bandpass function also depends on the carrier frequency with less roughness at lower and higher carrier frequencies than mid-carrier frequencies. These bandpass functions are analogous to the bandpass filters core to the modulation filter-bank model of Dau *et al.* (1997) that has been used to model auditory temporal processing in a variety of perceptual tasks.

Fastl and Zwicker (2007) also reported the dependence of perceived roughness on modulation depth using a tonal carrier (1000 Hz) amplitude modulated at a rate of 70 Hz. The modulation depth varied in a series of listening conditions from $m = 0.2$ to $m = 1.0$, spanning a range from relatively moderate roughness (i.e., somewhat smooth) to very rough. They showed that the magnitude of perceived roughness is a linear function of modulation depth when expressed on log–log axes (i.e., stimulus modulation depth in dB versus perceived roughness on a $\log_{10}$ scale) and is well represented by a power function with an exponent of about 1.6 for tonal carriers and of 1.3 for noise carriers.

FIG. 1. Roughness perception (in aspers) as a function of amplitude modulation frequency for tonal carriers of 125, 1000, and 8000 Hz. Figure adapted from Zwicker and Fastl (1990).

To better understand the perceived roughness of vocal stimuli, including dysphonic voices that exhibit abnormal roughness due to pathology, the current investigation is designed to determine the degree to which the perceived roughness associated with speech and speech-like stimuli follows the key results of the seminal research by Fastl and Zwicker (2007) described above. Specifically, one goal is to determine the dependence of the perceived vocal roughness on modulation frequency. A second goal is to determine the dependence of the perceived vocal roughness on modulation depth. If the perception of roughness for synthetic voice stimuli follows the same rules that apply to tonal and noise stimuli, then future work may be able to leverage quantitative models of roughness developed for non-speech stimuli to predict the perceived roughness of dysphonic voices.

II. EXPERIMENT 1: THE RELATIONSHIP BETWEEN PERCEIVED VOCAL ROUGHNESS AND AMPLITUDE MODULATION FREQUENCY (f_{AM})

If the perception of vocal roughness is governed by the same perceptual rules that govern the roughness of non-speech sounds, then the dependence of perceived vocal roughness on amplitude modulation frequency should follow the general form of the function shown in Fig. 1 for tonal stimuli.

A. Subjects

Subjects included ten (eight female, two male) listeners with normal hearing ranging in age from 18 to 26 yr (mean 20.9 yr, standard deviation 2.1 yr). Normal hearing was defined by pure tone thresholds < 20 dB hearing level (HL) at octave intervals from 250 to 8000 Hz (ANSI, 2010). Subjects were not required to have any clinical expertise in the evaluation of voice pathology. Participation was voluntary and all subjects consented to participate in the procedures as described below and approved by the Institutional Review Board. Subjects were paid for their participation.

B. Standard stimuli

To allow precise control of the roughness dimension, natural speech stimuli were selected, analyzed, re-synthesized with minimal amplitude fluctuation, and then amplitude modulated with a range of modulation depths (chosen to span a wide range of perceived roughness) and modulation frequencies (chosen with respect to data from Fastl and Zwicker, 2007). Stimuli were based on the /a/ vowels spoken by four different talkers taken from the Sataloff Heman-Ackah disordered voice database. These four stimuli were chosen because in their natural form they represented a wide range of perceived roughness (based on a previous study by Eddins *et al.*, 2013) and a wide range of fundamental frequency ($f_0 = 130, 148, 175$, and 252 Hz). The stimuli were re-synthesized based on the fundamental frequency and first five formants of the original vowels using a Klatt synthesizer (Klatt and Klatt, 1990). To parametrically control perceived roughness, the natural variation in amplitude was not included in re-synthesis. Rather, following synthesis, amplitude

fluctuation was controlled by superimposing sinusoidal amplitude modulation with modulation frequencies of either 10, 15, 20, 30, 40, 50, 70, or 90 Hz and a modulation depth of 100% or $m = 1.0$. Amplitude modulation is defined as

$$/a/_{\text{am}}(t) = /a/ \cdot [1 + m \cdot \sin(2 \cdot \pi \cdot f_{\text{am}} \cdot t)], \quad (1)$$

where $/a/$ is one of the four synthetic $/a/$ vowels, m is modulation depth (ranging from 0 to 1.0), f_{am} is the frequency of amplitude modulation in Hz, and t is time in seconds. The modulation depth is typically reported in logarithmic units such that the depth of AM, d_{AM} (in dB) = $20 \log_{10}(m)$.

C. Comparison stimuli

The comparison stimulus to be manipulated and matched to the standard stimuli is the same stimulus used in previous roughness matching investigations (e.g., Patel *et al.*, 2012). A sawtooth noise (fundamental frequency of 151 Hz) was chosen to have a vowel-like quality, spectrum, and waveform and was low-pass filtered (151 Hz cutoff frequency, -12 dB/octave) to match the average spectrum of several dozen dysphonic voices from the Sataloff Heman-Ackah database. To provide a more natural speech-like sound and to support breathiness matching (in other experiments), the filtered sawtooth wave was mixed with noise having the same spectral shape, resulting in a 20 dB signal-to-noise ratio (SNR). As described by Patel *et al.* (2012), the fluctuation needed to span the range of roughness from extremely smooth to extremely rough (dysphonic) was achieved by amplitude modulating the sawtooth plus noise waveform by a sinusoidal envelope raised to the fourth power. The independent variable in the matching task was the depth of modulation, d_{AM} (in dB). Thus, the perceived roughness of each of the standard stimuli was indexed by the modulation depth of the comparison stimulus.

D. Psychophysics

Perceived roughness was measured using a classical matching task following the methods of Patel *et al.* (2012). To estimate the perceived roughness of the voice stimulus of interest (the standard stimulus), the modulation depth of an amplitude modulated sawtooth wave mixed with noise (the comparison sound) was adjusted until the listeners judged the standard and comparison sounds to be equal in perceived fluctuation. Thus, on each trial, two sounds were presented—the standard followed by the comparison. In a graphical user interface (GUI), listeners were presented with three buttons labeled “increase fluctuation,” “decrease fluctuation,” and “equal fluctuation.” Accordingly, the modulation depth of the comparison sound was increased or decreased until the fluctuation of the comparison matched that of the standard sound. A series of trials leading to a judgment of equal fluctuation was considered a “run” and a block of trials consisted of five such runs. Other than the use of the term fluctuation above, there was no attempt to define the concept of roughness. To account for hysteresis (e.g., Gescheider, 1976), listeners completed one block of trials consisting of five descending runs that began with a very

high modulation depth (-10 dB modulation depth for the 10 and 90 Hz modulation frequencies and -5 dB modulation depth for the remaining modulation frequencies) and a second block of trials consisting of five ascending runs that began with a very low modulation depth (-25 dB) as described by Patel *et al.* (2012). The final matching threshold was based on the average matched threshold of the five ascending and five descending runs. The depth corresponding to matching threshold is expressed as modulation depth in dB or $20 \log_{10}(m)$, where m is the modulation index ($0 < m \leq 1.0$). Subjects completed the experiment in up to five sessions that were two hours or less in duration.

E. Test environment

Testing was completed in a double-walled, sound-attenuating chamber. Stimulus presentation and response collection was controlled by the TDT SykofizX software application (Tucker-Davis Technologies, Alachua, FL). Stimuli were routed from personal computer to a TDT RP2 real-time processor, TDT PA5 programmable attenuator, and TDT HB7 headphone buffer to a single (right) Etymotic (Elk Grove Village, IL) ER-2 insert earphone. Earphone output was calibrated to 75 dB SPL. Monaural presentation was chosen because modeling of voice quality often depends on loudness models that do not easily account for binaural interaction effects. Subjects were seated in front of a LCD monitor that displayed a GUI used for response collection. Buttons on the display allowed the independent variable to be increased, decreased, or to indicate a match.

F. Results

To examine the dependence of perceived roughness on modulation frequency, roughness matches (in dB) are shown as a function of modulation frequency in Fig. 2(a) as separate symbols for each base voice as labeled by their respective fundamental frequency. As can be seen in the legend inset, the four voice stimuli differed from each other in fundamental frequency ($f_0 = 130, 148, 175$, and 252 Hz). To determine the potential dependence of perceived roughness on modulation frequency and fundamental frequency, a two-factor repeated measures analysis of variance (ANOVA) was completed. The effect of modulation frequency was significant ($F_{7,3} = 9.50$, $p = 0.046$; partial $\eta^2 = 0.96$) and *post hoc* analyses indicated that matching for 10 Hz differed significantly from 15, 20, and 30 Hz, while matching for 90 Hz differed significantly from 30 and 40 Hz. There was no significant effect of fundamental frequency ($F_{3,7} = 1.85$, $p = 0.226$, partial $\eta^2 = 0.44$). Potential interaction between the two factors could not be evaluated due to insufficient residual degrees of freedom. Given the lack of an f_0 effect, the matching thresholds were averaged across the four voices as shown by the solid line and symbols in Fig. 2(b). Perceived roughness of voiced stimuli increases sharply with increasing modulation frequency from 10 to about 20 Hz, where perceived roughness shows a broad peak from about 20 to 50 Hz and declines gradually above 50 Hz modulation frequency.

FIG. 2. (Color online) Perceived roughness as a function of amplitude modulation frequency for voice stimuli. (a) Roughness for each of four different voices labeled by f_0 . (b) Roughness judgments averaged across the four voices.

III. EXPERIMENT 2: PERCEIVED VOCAL ROUGHNESS AS A FUNCTION OF AMPLITUDE MODULATION DEPTH (d_{AM})

If the perception of vocal roughness is governed by the same perceptual rules that govern the roughness of non-speech sounds, then the dependence of perceived vocal roughness on amplitude modulation depth should follow a linear function when modulation depth and roughness are expressed on logarithmic axes, parallel to the data reported by Fastl and Zwicker (2007) for tonal stimuli.

A. Subjects

Subjects included ten females with normal hearing ranging in age from 19 to 23 yr (mean 20.4 yr, standard deviation 1.1 yr). One of the ten subjects also participated in experiment 1. All subjects met the hearing sensitivity and consenting criteria as described in experiment 1.

B. Stimuli

Two of the four synthetic stimuli used in experiment 1 were selected as base stimuli for this experiment representing one male ($f_0 = 148$ Hz) and one female ($f_0 = 252$ Hz). From these 2 stimuli, 30 standard stimuli ($2 \text{ talkers} \times 5d_{AM} \times 3f_{AM}$) were created varying in modulation depth ($d_{AM} = 0, -5, -10, -15, \text{ and } -20$ dB) and modulation frequency

($f_{AM} = 20, 30, \text{ and } 45$ Hz). These five depths span a wide range of perceived roughness, and the three modulation frequencies span the pass-band of the curve shown in Fig. 1. The comparison stimulus was identical to experiment 1.

C. Psychophysics and experimental setup

The matching task and experimental setup were identical to experiment 1. Subjects completed testing in 3–4 sessions lasting between 1 and 1.5 h each.

D. Results

To determine the relationship between roughness superimposed upon the base voiced stimuli and the factors of modulation frequency and talker differences (male/female; $f_0 = 148$ or 252 Hz), a three-way repeated measures ANOVA was computed with modulation depth, modulation frequency, and talker (f_0) as within-subject variables. There was a significant effect of modulation depth ($F_{4,6} = 32.938, p < 0.001$, partial $\eta^2 = 0.956$). *Post hoc* analyses indicated a significant difference between all condition pairs ($p < 0.001$). The factors of modulation frequency ($F_{2,8} = 0.130, p = 0.880$, partial $\eta^2 = 0.031$) and talker/ f_0 ($F_{1,9} = 0.47, p = 0.834$, partial $\eta^2 = 0.005$) were not significant and there was no significant interaction among factors.

These results are easily visualized in Fig. 3. In Fig. 3(a), roughness matching values are averaged across the ten subjects and two talkers and are shown as a function of modulation depth with modulation frequency indicated by the different open symbols and dashed lines. There is a near-linear relation between perceived roughness and modulation depth. In Fig. 3(b), roughness matching values are plotted separately for the female and male voice and averaged across modulation frequency. Thus, over the range of f_0 values explored here (130–252 Hz), there is little dependence of roughness on f_0 . The data averaged across all four talkers were fit well with a linear function ($r^2 = 0.98$), although there was a slight deviation at the lowest modulation depth (-20 dB) reflecting slightly more perceived roughness for the male than female talkers.

IV. DISCUSSION

The primary goal of this study was to determine the similarity, if any, of perceived roughness in voice stimuli as compared to perceived roughness of tonal stimuli. The perceived roughness of tonal stimuli serves as the basis for models of both roughness perception and auditory temporal processing. For the voice stimuli shown here, the perceived roughness as gauged by a single-variable roughness matching task was strongly dependent on modulation frequency. The form of the function described by the matched threshold to modulation frequency has a similar bandpass shape characteristic of the roughness curves described by Fastl and Zwicker (2007) for tonal stimuli. Thus, the data from experiment 1 indicate that perceived roughness is similar for dysphonic voices and non-speech sounds. In addition to being similar in form, the Fastl and Zwicker (2007) data set shown in Fig. 1 (from their Fig. 11.2) indicate a strong dependence

FIG. 3. (Color online) Perceived roughness as a function of amplitude modulation depth for voice stimuli. (a) Roughness perception as a function of amplitude modulation depth is shown for each of three different modulation frequencies (fAM = 20, 30, and 45 Hz), indicated by separate symbols and lines. Each symbol represents the average across two voice stimuli ($f_0 = 148$ and $f_0 = 252$ Hz). (b) The same data are replotted to emphasize the potential influence of fundamental frequency. The two symbol types indicate fundamental frequency (male = 148 Hz, female = 252 Hz), while the data averaged across modulation frequencies of 20, 30, and 45 Hz.

of perceived roughness on carrier frequency (from 125 to 8000 Hz). The voiced stimuli in the current study can be considered complex tonal carriers with an average (across the four voices) f_0 of 176 Hz. Detailed comparison indicates that roughness in the current voiced stimuli [average data in Fig. 3(b)] is most similar to roughness for the 125 Hz carrier from Fastl and Zwicker (2007) in terms of the curve center frequency, bandwidth, and dynamic range.

Presumably, if roughness is related to amplitude modulation, then roughness perception should depend on the modulation depth, as well as modulation frequency. Experiment 2 demonstrated a nearly linear relationship between the depth of modulation in the voiced stimuli and its perceived roughness. Linear regression for each curve in Fig. 3 resulted in an average slope of 0.94 with an $r^2 = 0.99$. Fastl and Zwicker (2007) also investigated the dependence of perceived roughness on modulation depth using a 70 Hz modulation

rate and a 1000 Hz carrier stimulus presented at 60 dB SPL. When plotted on a log-log scale (stimulus modulation depth in dB versus perceived roughness), Fastl and Zwicker reported a near linear relationship with a slope of 1.6 (1.3 for broadband noise carriers). The data in Fig. 3(a) are well described by a similar function with an exponent of 0.96 ($R^2 = 0.98$). It is likely that the exponent (slope) depends somewhat on factors such as carrier type, large variations in carrier frequency, and, perhaps, carrier level.

Importantly, having demonstrated the robust similarity of roughness perception for the vocal stimuli used here and for pure-tone and broadband noise stimuli from Fastl and Zwicker, one would presume that computational models that have been used to predict perceived roughness of non-voice stimuli may be extended to predict the perceived roughness of voice stimuli. Such an application of general quantitative methods may prove useful in the quantification of dysphonic voices characterized by abnormal perceived roughness.

V. CONCLUSIONS

The influence of modulation frequency and modulation depth on the perceived roughness of voiced stimuli was measured in a group of normal-hearing subjects. These data demonstrate a strong dependence of roughness perception on modulation frequency and on modulation depth. Furthermore, this dependence closely matches roughness perception measures in previous psychoacoustic studies using tone and broadband noise carriers (e.g., Fastl and Zwicker, 2007). For tonal stimuli, they demonstrated that perceived roughness is greatest for mid-modulation frequencies and is less for lower and higher modulation frequencies, and that the mid-frequency is dependent on the carrier frequency. This dependence on modulation frequency is similar for voice stimuli, and the form of the function relating modulation frequency to roughness perception for voices with an average $f_0 = 176$ Hz best maps to the function for a tonal carrier of 125 Hz. While the current data do not show a significant dependence of roughness on f_0 , the average f_0 for pediatric voices is much higher (e.g., 330 Hz). Based on previous data for tonal carriers, roughness might peak at a modulation frequency closer to 60 Hz. These relationships set the stage for the extension of models of roughness perception and amplitude modulation perception to the perceived roughness of dysphonic voices. To the extent that such models can accurately predict perceived roughness associated with dysphonic voices, one may be able to realize objective measures of voice quality that could have great clinical and theoretical significance.

ACKNOWLEDGMENTS

This work was supported by National Institutes of Health National Institute on Deafness and Other Communication Disorders (NIH NIDCD) Grant No. R01 DC009029.

¹This work originally appeared in the first edition by Zwicker and Fastl (1990).

- Dau, T., Kollmeier, B., and Kohlrausch, A. (1997). "Modeling auditory processing of amplitude modulation. I. Modulation detection and masking with narrowband carriers," *J. Acoust. Soc. Am.* **102**, 2892–2905.
- Eddins, D. A., and Shrivastav, R. (2013). "Psychometric properties associated with perceived vocal roughness using a matching task," *J. Acoust. Soc. Am.* **134**(4), EL294–EL300.
- Fastl, H., and Zwicker, E. (2007). *Psychoacoustics: Facts and Models*, 3rd ed. (Springer, New York), pp. 257–264.
- Gescheider, G. A. (1976). *Psychophysics: Method and Theory* (Erlbaum, Hillsdale, NJ), pp. 45–72.
- Hirano, M. (1981). *Clinical Examination of Voice* (Springer, New York), p. 81.
- Kempster, G. B., Gerratt, B. R., Verdolini-Abbott, K., Barkmeier-Kraemer, J., and Hillman, R. E. (2009). "Consensus auditory-perceptual evaluation of voice: Development of a standardized clinical protocol," *Am. J. Speech Lang. Pathol.* **18**, 124–132.
- Klatt, D. H., and Klatt, L. C. (1990). "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Am.* **87**, 820–857.
- Patel, S., Shrivastav, R., and Eddins, D. A. (2012). "Identifying a comparison for matching rough voice quality," *J. Speech Lang. Hear. Res.* **55**, 1407–1422.
- Terhardt, E. (1974). "On the perception of periodic sound fluctuations (roughness)," *Acustica* **30**(4), 201–213.
- Zwicker, E., and Fastl, H. (1990). *Psychoacoustics: Facts and Models* (Springer, New York), pp. 1–354.
- Zwicker, E., and Fastl, H. (1999). *Psychoacoustics: Facts and Models*, 2nd ed. (Springer, New York), pp. 1–428.