Research Article

Predicting Perceived Vocal Roughness Using a Bio-Inspired Computational Model of Auditory Temporal Envelope Processing

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

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

ARTICLE INFO

Article History:
Received February 15, 2022
Revision received April 14, 2022
Accepted April 25, 2022

Editor-in-Chief: Susan L. Thibeault

https://doi.org/10.1044/2022_JSLHR-22-00101

ABSTRACT

Purpose: Vocal roughness is often present in many voice disorders but the assessment of roughness mainly depends on the subjective auditory-perceptual evaluation and lacks acoustic correlates. This study aimed to apply the concept of roughness in general sound quality perception to vocal roughness assessment and to characterize the relationship between vocal roughness and temporal envelop fluctuation measures obtained from an auditory model.
Method: Ten /ɑ/ recordings with a wide range of roughness were selected from an existing database. Ten listeners rated the roughness of the recordings in a single-variable matching task. Temporal envelope fluctuations of the recordings were analyzed with an auditory processing model of amplitude modulation that utilizes a modulation filterbank of different modulation frequencies. Pitch strength and the smoothed cepstral peak prominence were also obtained for comparison.
Results: Individual simple regression models yielded envelope standard deviation from a modulation filter with a low center frequency (64.3 Hz) as a statistically significant predictor of vocal roughness with a strong coefficient of determination (r^2 = .80). Pitch strength and CPPS were not significant predictors of roughness.
Conclusion: This result supports the possible utility of envelope fluctuation measures from an auditory model as objective correlates of vocal roughness.

Voice disorders typically lead to perceptual changes in voice known as dysphonia. As such, assessment of voice quality is an integral component of a complete and accurate evaluation of dysphonia and repeated assessments of voice quality often serve as essential outcome measures for the treatment of voice disorders (Behrman, 2005; Carding et al., 2009). The most common assessment of voice quality is a subjective auditory-perceptual evaluation performed by a speech-language pathologist specializing in clinical voice assessment. Components of the perceptual voice assessment include overall dysphonia severity and parsing overall dysphonia into primary voice quality dimensions of breathiness, roughness, and strain as well as other dimensions such as nasality, pitch, and loudness (Hirano, 1981; Kempster et al., 2009). Research has sought to develop objective acoustic measures that can quantify voice quality with the hopes of improving accuracy, reliability, and efficiency of assessment and outcomes measures. One such measure is the cepstral peak prominence (CPP; Heman-Ackah et al., 2003), which that serves as a constituent component of multivariate acoustic models such as Acoustic Voice Quality Index (Maryn et al., 2010) and Cepstral Spectral Index of Dysphonia (Awan et al., 2013). These multivariate acoustic models yield a strong correlation with perceptual ratings of overall dysphonia severity (Awan et al., 2013; Maryn et al., 2010). However, strong acoustic predictors of breathy, rough, and strain voice qualities are still lacking.

Correspondence to Yeonggwang Park: park21@usf.edu. ***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

Vocal roughness is manifest in various voice pathologies (e.g., vocal fold nodules, polyps, and paralysis) and is typically defined as perceived irregularity in vocal fold vibration (Hirano, 1981; Kempster et al., 2009). Because of the irregularity perceived in rough voices, cycle-by-cycle fluctuations of frequency and amplitude such as jitter and shimmer have been investigated as possible acoustic correlates of perceived roughness; however, the relationship between perturbation measures and perceived roughness was found to be mostly weak to moderate ($|r|$ = .29–.71; Barsties v. Latoszek, Maryn, et al., 2018; Bhuta et al., 2004). Perturbation measures require accurate fundamental frequency (f_o) estimation, which is difficult for many dysphonic voices (Mehta & Hillman, 2008), and as such, they are no longer recommended as a clinical measure of voice quality (Patel et al., 2018). Spectral noise measures and CPP also have been investigated as possible correlates of roughness, but their relationships with roughness often are also weak ($|r|$ = .02–.43; Barsties v. Latoszek, De Bodt, et al., 2018; Heman-Ackah et al., 2002). In an effort to separate breathy from rough voices in dysphonic and normophonic adults, Awan and Awan (2020) used a variation in the CPP known as the $\text{CPP}_{\text{High-Low}}$ with which they considered CPP computed across high- versus low-quefrency portions of the cepstrum. That measure was successful at classifying the voices into separate qualities. However, no measure to date has accurately captured roughness over the range considered "normal" to "extremely rough."

To develop a better objective correlate of vocal roughness, here, we consider the possibility that the acoustic correlate of roughness is analogous to a change in the temporal envelope of a waveform. The auditory percept of roughness has been associated with both frequency modulation and amplitude modulation. Indeed, when the range of frequency modulation exceeds upper or lower filter cutoff, the resulting output contains amplitude modulation. The linkage between auditory perception of roughness and amplitude modulation frequency and depth was summarized by Fastl and Zwicker (2007). They investigated amplitude modulation of pure tones and broadband noise and showed that the roughness percept was associated with relatively low modulation frequencies, ranging from 15 to 300 Hz. For pure tones of 125 and 250 Hz, similar to the human vocal f_o, perceived roughness was maximum when the modulation frequencies were between approximately 25 and 50 Hz. For a 1000 Hz tone and broadband noise, perceived roughness was maximum for a modulation frequency of approximately 70 Hz. In a systematic investigation of the relationship between amplitude modulation and vocal roughness, we applied amplitude modulation to voices that were judged by expert listeners to have no perceived roughness. Following amplitude modulation, the voices were judged to have maximum perceived roughness when the frequency of the applied amplitude modulation was between 20 and 50 Hz (Eddins et al., 2015). Similarly, applying amplitude modulation to a synthetic speech-shaped waveform, and varying the amplitude modulation depth, has served well as a comparison sound in a single-variable matching task for vocal roughness evaluation (Patel et al., 2012). Each of these results leads to the postulation that measures that quantify the temporal envelope of an acoustic stimulus spanning a range of amplitude modulation frequencies may provide useful predictions of vocal roughness.

While acoustic correlates of voice quality may be elusive, Shrivastav and colleagues (Shrivastav, 2003; Shrivastav & Sapienza, 2003) pointed out that objective predictors of the auditory perception of voice quality may require nonlinear transformations analogous to those that occur in the auditory system as the sound is transformed from an acoustic waveform that enters the outer ear to a neural code that leads to a given perception. Accordingly, computational models of auditory processing have been used to process the vocal acoustic signal and the outputs of such models have resulted in strong correlations with auditory-perceptual evaluation of vocal attributes. These include a bio-inspired model of pitch strength that correlates strongly with breathiness (Eddins et al., 2016) and a bio-inspired model of the auditory perception of sharpness that correlates strongly with perceived vocal strain (Anand et al., 2019). Here, we hypothesized that a bio-inspired model of auditory temporal envelope processing may be used to accurately predict the perception of vocal roughness.

The temporal modulation filter bank model first described by Dau et al. (1997a), revised several times since, is a strong candidate for predicting vocal roughness as it has produced accurate predictions of the most comprehensive set of auditory-perceptual phenomena to date. In this report, we have adapted the most recent version of that model (Majdak et al., 2021) in an effort to establish an objective correlate to perceived vocal roughness. Briefly, the model includes a filter bank in the audio frequency domain that mimics cochlear tonotopicity. This filtering is followed by rectification analogous to hair-cell transduction and nonlinear adaptation mirroring physiological damping. Subsequently, the rectified and damped output of each audio-frequency filter is processed by a second filter bank in the temporal modulation domain that reflects the ensemble characteristics of auditory midbrain tuning to temporal envelopes. Model output is analogous to the internal representation of the temporal envelope within the central auditory system.

The purpose of the current investigation is to evaluate the ability of an auditory temporal modulation filter bank model to predict the auditory perception of roughness in a set of sustained vowels that were selected to be primarily rough with minimal breathiness and strain. Perception was measured using a single-variable matching task (i.e., Patel et al., 2012), model output was computed, and model output and perceptual judgments were compared.

Method

Stimuli

Voice Stimuli

Sustained /ɑ/ recordings from 10 individuals (seven women and three men; M_{age} = 59.9 years, range: 27–85 years) spanning a wide range of roughness (from least to most roughness) were selected using stratified random sampling (Shrivastav et al., 2005) from the University of Florida Dysphonic Voice Database. The individuals had various voice disorders including vocal nodules, paralysis, laryngeal cancer, dystonia (abductor), granuloma, papilloma, presbyphonia, muscle tension dysphonia, and glottic and subglottic stenosis. One individual had a thyroid tumor. Selected recordings had stable f_o contours and were primarily rough with minimal breathiness or strain. These criteria were chosen to increase the probability that perceptual judgments could be limited to the roughness percept. From the original recordings, the central 500-ms portion was excised and downsampled to 24414 Hz to match the hardware requirements for the perceptual experiment.

Comparison Sounds

The comparison sound was used as the comparison to the natural /ɑ/ vowel samples in the matching task that was designed to estimate perceived roughness. The comparison sound consisted of a sawtooth waveform mixed with noise. Prior to mixing, both stimuli were low-pass filtered to have a spectral tilt of −12 dB/octave. The combination, sawtooth + noise, had a sawtooth-to-noise ratio of 20 dB, which achieved a more natural speech-like quality. The f_o of the sawtooth was set to 151 Hz. The filter slope and f_o were based on the average spectral slope and f_o of a set of dysphonic voices in the Massachusetts Eyes and Ear Infirmary (MEEI) Disordered Voice Database (MEEI Voice and Speech Laboratory, 1994). To introduce amplitude fluctuation, which is perceived as roughness (Eddins et al., 2015; Fastl & Zwicker, 2007), the sawtooth-plus-noise complex was amplitude modulated during the matching task with a sinusoidal function having a frequency of 25 Hz and raised to the fourth power. This produced a waveform shape with relatively sharp peaks and broad valleys. By altering the modulation depth, m, expressed in dB as 20 $\log_{10}(m)$, this comparison sound has been shown to produce a range of perceived roughness values wide enough to exceed the range required to match dysphonic voices (Eddins & Shrivastav, 2013). Following amplitude modulation, the root-mean-square (RMS) amplitude of each comparison sound waveform was normalized to a constant RMS value.

Single-Variable Matching Task

The auditory-perceptual task chosen for this experiment was a single-variable matching task where the single independent variable parameter was amplitude modulation depth. On each trial of the matching task, listeners heard the target sound (voice sample) followed by a 500-ms silent interval and then by a comparison sound (synthetic waveform described above). Simultaneous with sound presentation, a graphical user interface (GUI) was displayed on the computer monitor in front of the listener. The GUI included three buttons labeled “increase fluctuation,” “decrease fluctuation,” and “equal fluctuation.” When listeners perceived the roughness of the comparison sound to be less than the perceived roughness of the target sound, they selected the “increase fluctuation” button via mouse click. This resulted in an increased amplitude modulation depth of the comparison sound on the next trial. When they perceived the roughness of the comparison sound to be greater than the perceived roughness of the target sound, they selected the “decrease fluctuation” button. This resulted in a decreased amplitude modulation depth on the next trial. When the point of subjective equality was reached, they selected the “equal fluctuation” button. Prior to testing, listeners were instructed to focus only on the fluctuation or roughness of each sound and to ignore other percepts such as pitch, loudness, timbre, breathiness, strain, and vowel identity.

For each voice stimulus, each subject produced 10 perceptual matches based on five descending series and five ascending series of trials as described by Patel et al. (2012). For descending series, the initial modulation depth was −5 dB, which had a perceived fluctuation that exceeded any dysphonic voice, resulting in an initial response of “decrease fluctuation.” For ascending series, the initial modulation depth was −30 dB, which had a perceived fluctuation that was less than any dysphonic voice, resulting in an initial response of “increase fluctuation.” Following each response, the amplitude modulation depth was decreased or increased in 2-dB steps until a perceptual match was indicated. By averaging the five descending and five ascending runs, hysteresis associated with the adaptive tracks was averaged out and a stable matching threshold (roughness matching value) was obtained.

The matching task was chosen because, compared with other perceptual tasks, such as visual analog scales or magnitude estimation, a matching task can minimize the effects of context and internal biases and provides more reliable perceptual measurement (Kreiman & Gerratt, 1998, 2005; Patel et al., 2010). The reliability of the perceptual measure was particularly important in this experiment because perceptual measures were used to evaluate the validity of possible acoustic correlates or model predictions of perceived vocal roughness.

Instrumentation

The Tucker-Davis Technologies (TDT) SykofizX software application was used to load from disk target voice samples from stored audio files and to compute the required comparison sound for each trial. The software controlled the stimulus presentation, the subject interface, and response collection. A TDT real-time processor (Model RZ6) converted the digital files to analog signals that were routed to a programmable attenuator (TDT PA5) and headphone buffer (TDT HB6) and delivered to an Etymotic ER-2 insert earphone at a calibrated level of 85 dB SPL.[1]

Listeners

Fifteen participants (14 women and one man) ranging in age from 19 to 37 years (M_{age} = 23.7 years) were recruited as listeners. All listeners had pure-tone thresholds less than 25 dB HL at frequencies of 125, 250, 500, 1000, 2000, 4000, and 8000 Hz (ANSI, 2010). Listeners were native speakers of American English and had no previous training in voice quality evaluation. All listeners consented to participate according to procedures approved by the University of South Florida biomedical institutional review board (Protocol Pro00012381) prior to engaging in study activities and were paid an hourly rate ($12 USD) for their participation.

Task Familiarization

Prior to performing the matching experiment, listeners were first familiarized with the matching task using synthetic target sounds identical to the comparison sound with fixed modulation depths of either 0, −5, −10, or −15 dB. The order of the target stimuli was pseudorandomized for each listener, but the stimulus with the modulation depth of −15 dB, which contained relatively little modulation depth, was not placed first for any listener in order to reduce confusion in the first few trials. During task familiarization, the experimenter provided verbal feedback and additional instruction if listeners had difficulty matching the comparison sound to the target stimulus.

After completing the familiarization task with four synthetic stimuli, listeners performed additional practice tasks with two natural vowel /ɑ/ samples that were not part of the 10 experimental stimuli. This step ensured that listeners were able to perform the matching task appropriately with natural voices. One sample was rough, and the other was less rough.

Experimental Task

The experimental matching task was performed with the 10 natural /ɑ/ stimuli. The order of the 10 stimuli was pseudorandomized for each listener via MATLAB (The MathWorks, Inc) algorithm. Listeners completed the matching task for two target stimuli at a time followed by a short break. The order of the stimuli and initial independent variable value also was pseudorandomized within a set by SykofizX software. Each set took approximately 15–20 min. Listeners completed data collection in two to three sessions lasting no more than 2 hr including breaks.

Auditory Temporal Modulation Filter Bank Model Processing

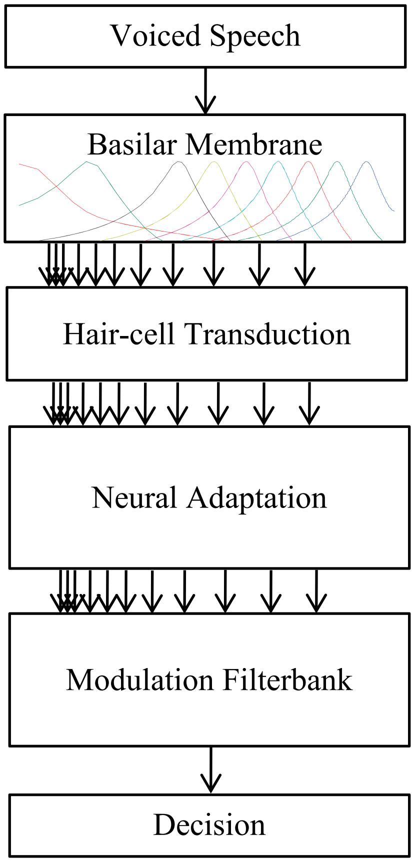
The auditory filter bank model chosen for use in this study was described by Dau et al. (1997a) and was chosen because of the ability of the model to predict a wide variety of data related to the perception of the temporal envelope of an acoustic stimulus. The model also has been used to predict a variety of simultaneous and masking experiments (e.g., Dau et al., 1997b) as well as other auditory-perceptual abilities. This simple model requires no optimization or training on any preliminary data sets. As shown in Table 1, the model consists of the following processing steps: (a) a basilar membrane stage, consisting of a linear gammatone filterbank (Patterson et al., 2003) to separate audio frequency bands for subsequent processing; (b) a hair-cell transduction stage, modeled by a half-wave rectifier and low-pass filter at 1000 Hz per audio frequency channel; (c) an auditory nerve stage, modeled by frequency-dependent nonlinear adaptation (Münkner, 1993); and (d) a modulation filterbank to account for sensitivity to

[1] Each digital vowel stimulus waveform (500 ms) was digitally scaled prior to digital-to-analog (D/A) conversion, setting the RMS level to −6 dB FS. Prior to level calibration, each vowel waveform was duplicated and concatenated to extend the duration to 4 s. Analog stimuli were subsequently attenuated by a vowel-specific attenuation value using an external digitally programmable attenuator (TDT PA5) to achieve a sound pressure level of 85 dB SPL measured as described below.

Comparison sound waveforms (500 ms) were digitally scaled prior to D/A conversion, setting the RMS level to −10.4 dB FS. Prior to level calibration, the waveform was duplicated and concatenated to extend the duration to 4 s. Analog stimuli were subsequently attenuated by a constant value, using an external digitally programmable attenuator (TDT PA5), to achieve a sound pressure level of 85 dB SPL measured as described below.

Sound level measurements involved connecting the ER-2 insert phone to its corresponding foam ear tip and inserting the eartip into a Zwislocki dB 100 ear simulator (Bruel & Kjaer) fit with a ½ pressure microphone (G.R.A.S., Model 40AG) connected to a preamplifier (G.R.A.S, Model 26AK). The preamp was routed to a power module (G.R.A.S., Model 12AA), the output of which was measured with a volt meter (Fluke, Model 45). The sound pressure level was determined relative to a reference voltage established with a calibrator (Bruel & Kjaer, Model 4230) connected to the ½-in. microphone, preamplifier, and power module circuit.

Table 1. Description of the processing steps of the auditory temporal modulation filter bank model.



Steps	Description
Voiced Speech	/ɑ/ recordings from University of Florida Dysphonic Voice Database
Basilar Membrane	Cochlear "gammatone" filter bank based on Patterson (2003): Center frequencies from 132 to 12207 Hz are uniformly spaced based on the ERB-rate scale (Moore & Glasberg, 1996).
Hair-cell Transduction	Low pass filter: First-order Butterworth with cutoff frequency of 1000 Hz, slope of −6 dB per octave; half-wave rectification: Values less than zero set to zero
Neural Adaptation	Cascade of five Butterworth first-order lowpass filters with time constants of 5, 50, 129, 253, 500 ms implemented as a feedback loop as described by Dau et al. (1996a)
Modulation Filterbank	Temporal modulation filter bank with 1 lowpass and 11 bandpass filters. Filter parameters are in Table 2. Implementation and envelope extraction are described in Dau et al. (1997a).
Decision	Standard deviation of envelope extracted from Modulation Filters 5, 6, or 7 (see text)

Note. Any model parameters not specified above were taken from Dau et al. (1997a). ERB = equivalent rectangular bandwidth.

amplitude changes in the temporal envelope of the acoustic waveform. Each stage of the auditory front end (Dau et al., 1997a) was programmed in MATLAB using open-source code from the auditory modeling toolbox version 1.0 (Majdak et al., 2021). In the current implementation, the gammatone filterbank (Step 1) consisted of 32 channels with center frequencies ranging from 132 to 12207 Hz spaced in equal steps of equivalent rectangular bandwidth (Moore and Glasberg, 1996). For any given noise or speech stimulus, the output of the peripheral model produces 12 arrays, one for each modulation filter (Step 4) with filter center frequencies ranging from ~5 Hz to 1000 Hz (see Table 2). After auditory processing of each stimulus waveform, the model returns a matrix of filter outputs as a function of time. Standard deviations of the outputs from Modulation Filters 5, 6, or 7 were computed and used for statistical analysis to test our primary hypothesis. Accordingly, these values will be denoted EnvSD_5, EnvSD_6, or EnvSD_7. Candidate Modulation Filters 5, 6, and 7 were chosen from among the full modulation filter bank because the center frequencies of these filters (23, 39, and 64 Hz) were in the range of the modulation frequencies that evoked greatest perceived roughness in human voice samples (Eddins et al., 2015) and correspond to the range amplitude modulation frequencies that resulted in the maximum roughness sensation in a broadband carrier (Fastl & Zwicker, 2007).

Other Acoustic Measures

In addition to the envelope standard deviations described above, we also obtained estimates of pitch strength and the CPPS from each voice sample to compare these additional acoustic measures with the envelope standard deviations from the auditory filter bank model described above. Pitch strength is the salience of pitch sensation (Shrivastav et al., 2012), and CPPS is the first cepstral peak amplitude normalized to the smoothed cepstral amplitudes of background noise (Hillenbrand et al., 1994;

Table 2. Center frequencies and bandwidths of the band pass filters that make up the auditory temporal modulation filter bank (Dau et al., 1997a).

Modulation filter	Center frequency (Hz)	Bandwidth (Hz)
1	N/A	2.5*
2	5	5
3	10	5
4	13.9	6.9
5	23.1	11.6
6	38.6	19.3
7	64.3	32.2
8	107.2	53.6
9	178.6	89.3
10	297.7	148.8
11	496.1	248.1
12	826.9	413.5

Note. N/A = not applicable.

*Filter 1 was a low-pass filter, so the bandwidth also denotes the filter cutoff frequency.

Maryn & Weenink, 2015). Pitch strength was estimated from a sawtooth waveform inspired pitch estimator with auditory front end (Camacho, 2012). CPPS was obtained using the built-in function in PRAAT (Boersma & Weenink, 2021) acoustic analysis software using the protocol described by Watts et al. (2017).

Statistical Analysis

Model Development

Intra- and interrater reliability of the matching task were estimated as intraclass correlation coefficient (ICC) in MATLAB. Simple linear regression models were computed in SPSS (Version 27, IBM Corp.) to evaluate the relationship between roughness matching values and the computational measures. The response variable in the models was the roughness matching value in dB amplitude modulation depth (dB d_{AM}). The predictor variables were the EnvSD_5, EnvSD_6, and EnvSD_7, pitch strength, and CPPS, and simple linear regression models were analyzed for each predictor variable. The significance level was adjusted to .01 with Bonferroni correction (p = .05/5 predictors = 0.01) to reduce the probability of a Type I error. Effect sizes of the significant predictors were estimated as Cohen's f^2.

Evaluation of Model Prediction Accuracy

We used the linear regression equation with the highest r^2 to evaluate the ability of the model to predict perceived roughness for novel data. A second set of 10 voices across a wide range of primary roughness was selected from the University of Florida Dysphonic Voice Database (five women and five men; M_{age} = 62 years; range: 47–73 years). Thirty new listeners (25 women and five men; M_{age} = 23.6 years; range: 19–37 years), who met the same criteria as the original listeners, were recruited and performed the same single-variable matching task described above. Perceived roughness of the 10 new voice stimuli was obtained from the matching task, and perceptual vocal roughness of the 10 new voice stimuli was predicted based on the linear regression equation obtained with the original data set. Pearson's r was computed in SPSS to evaluate the relationship between the perceived and predicted roughness.

Results

Perceived vocal roughness is shown in Figure 1 with voice sample on the abscissa and perceived roughness on the ordinate as estimated from the single-variable matching task and quantified in units of dB d_{AM}. Symbols reflect mean matching values for the 15 listeners with error bars indicating 95% confidence intervals. Stimuli are ordered in terms of perceived roughness from least to most rough. The range of matching values corresponds closely to those reported by Patel et al. (2012).

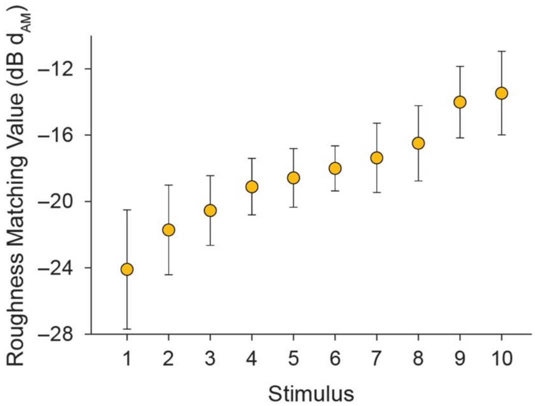
Listener Reliability

Intrarater reliability (ICC [2, k], absolute agreement) for the 15 listeners was high, ranging from .87 to .99 with a mean of .96. Interrater reliability (ICC [2, k], consistency) among the 15 listeners also was high with a value of .92.

Simple Linear Regression Models

Individual simple regression models for roughness matching values were computed for each computational

Figure 1. Mean roughness matching values for 10 stimuli across 15 listeners. Error bars indicate 95% confidence intervals.



measure and the temporal modulation filter bank model output. Table 3 presents r^2 and associated statistical values for each regression model. The effect size was calculated for each statistically significant predictor. The EnvSD_7 was a significant predictor of perceived vocal roughness, accounting for 80% (r^2 = .80) of the variance in the perceptual data and had a very large effect size (f^2 = 4.00). The EnvSD_5 and EnvSD_6 metrics were not significant predictor variables. Furthermore, subsequent to a log-transform of the EnvSD_7, $\log_{10}(\text{EnvSD}_7)$, a simple linear regression model with the log-transformed data accounted for 86% (r^2 = .86) of the variance in the perceptual data ($F_{1,8}$ = 49.6; p < .001; f^2 = 6.14). The regression equation of perceived roughness with $\log_{10}(\text{EnvSD}_7)$ as a predictor is described in Equation 1:

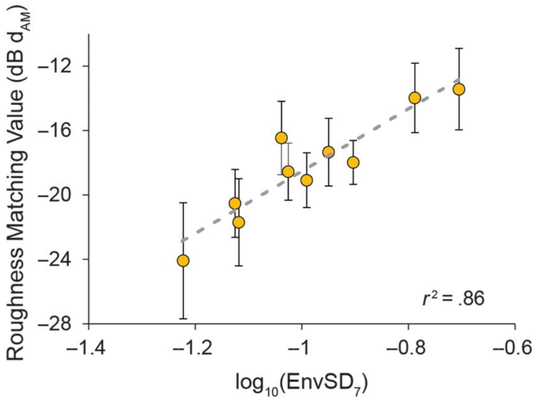
$$\text{Predicted Roughness}\,[\text{dB d}_{\text{AM}}] = 19.32 \times \log10(\text{EnvSD}_7) + 0.76 \qquad (1)$$

Figure 2 shows the relationship between the log-transformed EnvSD_7 metric and perceptual roughness matching values. As hypothesized, positive correlations indicated that stimuli with higher envelope standard deviation were perceived to have higher roughness than stimuli with lower envelope standard deviation. None of the other candidate acoustic metrics was a significant predictor of roughness matching values (each accounted for less than 50% of the variance in the perceptual data).

Example of Model Predictions

For this second listening experiment to evaluate model prediction accuracy, intrarater reliability (ICC [2, k], absolute agreement) ranged from .87 to .99, with a mean of .97, whereas interrater reliability (ICC [2, k], absolute agreement) was .98. Figure 3 displays perceived vocal roughness of the 10 new voice stimuli resulting from the matching task (dB d_{AM}) on the ordinate and predicted perceptual roughness from Equation 1 on the abscissa. Perceived roughness and predicted roughness were strongly and significantly correlated (r = .84, p = .001), accounting for approximately 71% (r^2 = .71) of the variance for this new set of stimuli. The roughness of most samples was well predicted by Equation 1 except for the two roughest samples (mean perceived roughness = −12.0 dB d_{AM} and − 13.6 dB d_{AM}) underestimated by the equation.

Figure 2. A scatter plot and a linear fit of mean roughness matching values as a function of $\log_{10}(\text{EnvSD}_7)$. Error bars indicate 95% confidence intervals.



Discussion

This study examined the relationship between perceived rough voice quality and envelope fluctuation measures obtained from a bio-inspired auditory model of temporal modulation processing. Perceived roughness of the 10 voice samples in a wide range of primary roughness was obtained from a single-variable matching task. Envelope standard deviations were computed from the output of specific modulation filters with low modulation frequencies (23 Hz–64 Hz) of the auditory model. The simple regression model of perceived roughness reported here, with a predictor based on the log-transformed EnvSD_7 resulted in a strong (r^2 = .86) coefficient of determination.

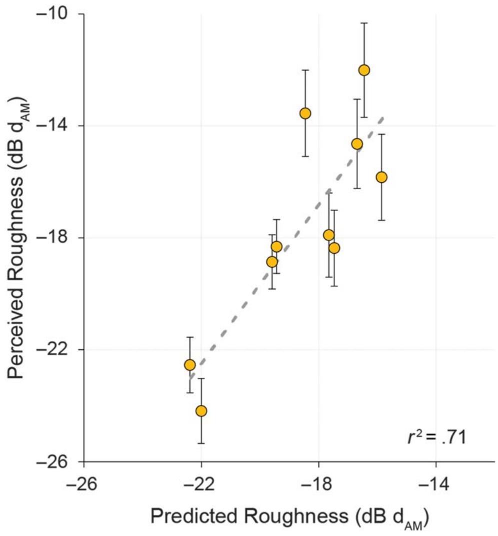
The results of this study are in line with previous studies that observed high correlations between voice quality perception and acoustic measures obtained from bio-inspired computational auditory models. Shrivastav and Sapienza (2003) reported that partial loudness estimates

Table 3. Coefficient of determination (r^2) and statistical values for each linear regression model.

Acoustic measure	r^2	$F_{1,8}$	Coef.	*SE* coef.	*t*	*p*	Effect size (f^2)
EnvSD_5	.36	4.47	295.26	139.71	2.11	.07	
EnvSD_6	.50	7.82	125.60	44.90	2.80	.02	
EnvSD_7	.80	30.94	68.98	12.46	5.56	**< .001**	4.00
Pitch strength	.32	3.82	−15.11	7.73	−1.96	.09	
CPPS	.07	0.59	−0.30	0.38	−0.77	.47	

Note. Bold value indicates significant results (p < .01). Coef. = coefficient; *SE* = standard error; CPPS = smoothed cepstral peak prominence.

Figure 3. A scatter plot and a linear fit of mean perceived roughness (roughness matching values) of the new 10 voice samples as a function of predictive roughness estimated with Equation 1. Error bars indicate 95% confidence intervals.



based on the periodic signal and noise computed by an auditory model front-end yielded a stepwise regression model of perceived breathiness that accounted for 85.2% of the variance. Eddins et al. (2016) also reported a strong linear regression model of perceived breathiness (r^2 = .87) using the pitch strength estimate obtained through a sawtooth waveform inspired pitch estimator with an auditory front-end (Camacho, 2012). In the case of strained voice quality, Anand et al. (2019) reported that spectral energy distribution measures obtained from an auditory model, combined CPP computed from the input waveform, resulted in a stepwise regression model of perceived strain that accounted for 77%–79% of the variance.

The current investigation, along with these previous studies, demonstrates that signal processing of raw acoustic signals using bio-inspired auditory processing front ends can be used to form models that effectively predict voice quality perception and thus highlight the potential for such measures to serve as objective measures of voice quality. The high correlations observed between voice quality perception and the output of bio-inspired models of auditory perception likely reflect the fact that such signal-processing front ends capture the essence of various nonlinear processing properties and transformations that occur in the peripheral auditory system during the auditory-perceptual process (Dau, 2008; Shrivastav & Sapienza, 2003). In contrast, conventional objective indices of vocal acoustic signals do not include those transformations. As a class, acoustic analyses of voice may be very useful measures for some purposes; however, it is not surprising that inclusion of processing steps related to the perception of sound may improve correspondence between objective measures and perceptual judgments. Such measures typically are more computationally expensive than conventional voice measures and are not commonly included in commercially available software.

Specifically, the auditory temporal envelope processing model used in this study contains a temporal modulation filter bank. This model extracts specific temporal envelope modulation frequencies at the output of various audio-frequency channels (Dau et al., 1997a), thereby approximating the internal representation of the temporal envelope by the auditory system, and in so doing appears to capture the primary characteristics that give rise to variations in perceived roughness magnitude. The modulation filter bank is analogous to the inherent tuning of single neurons and families of neurons in the central auditory system, from the cochlear nucleus up to the auditory cortex, to different amplitude modulation frequencies (e.g., Langner, 1992). Thus, each modulation filter represents an ensemble of cells that are tuned to the modulation frequencies passed by that filter. By having the modulation filter bank, the model can estimate the degree of envelope modulation at different modulation frequencies, which were observed to be useful in predicting vocal roughness in this study.

In the analyses reported here, the output from Modulation Filter Number 7 best predicted vocal roughness associated with the sustained vowel recordings in this study. The center frequency of this filter is 64 Hz with a bandwidth of 32 Hz. The modulation frequency range of this filter is similar to the frequency (~70 Hz) that resulted in the maximum perceived roughness associated with amplitude modulation superimposed on a broadband noise and 1000 Hz tone carriers, as reported by Fastl and Zwicker (2007). This range is slightly higher than the modulation frequency range (25–50 Hz) that was judged as most rough when amplitude modulation was applied to 125 and 250 Hz tones (Fastl & Zwicker, 2007) and natural adult normophonic voice samples (Eddins et al., 2015). One possible reason for this discrepancy in the modulation frequency range may be related to differences in the complexity of the temporal envelope, with more complex envelope fluctuations in natural rough voices than the simple envelope modulation function applied in previous studies. The sound samples of Fastl and Zwicker (2007) and Eddins et al. (2015) were applied with periodic amplitude modulation at a constant modulation frequency. Natural rough voices are likely to contain more random and irregular temporal fluctuation than periodic amplitude modulation applied in the previous studies. However, the current results agree with the results from previous studies in that roughness is associated with temporal envelope fluctuations at relatively low modulation frequencies.

Shimmer, a traditional measure of amplitude perturbation, was observed to have a weak to moderate correlation with perceived roughness in previous studies (Barsties v. Latoszek, De Bodt, et al., 2018; Bhuta et al., 2004). Shimmer estimates period-to-period amplitude perturbation. This period-to-period amplitude perturbation reflects the faster temporal fine structure of the waveform rather than the slower temporal envelope associated with the amplitude modulation frequencies of focus in this study. The fact that envelope frequencies in the 25–75 Hz range are more strongly related to roughness than higher envelope frequencies (Eddins et al., 2015; Fastl & Zwicker, 2007) may explain why shimmer has been not strongly correlated with perceived roughness previously. In addition, shimmer requires accurate estimation of f_o in order to be reliable, which is challenging in highly dysphonic voices (Mehta & Hillman, 2008).

Our results indicating that CPP and pitch strength were not significant predictors of perceived roughness are consistent with the notion that temporal envelope fluctuations may be a stronger constituent of perceived roughness than signal periodicity, a feature well-captured by CPP and pitch strength estimates. Both CPP and pitch strength are related to the degree of signal periodicity (Hillenbrand et al., 1994; Shrivastav et al., 2012) and were strongly correlated with perceived breathiness (Eddins et al., 2016; Hillenbrand et al., 1994). Aperiodicity of the signal has been also associated with perceived roughness (Barsties v. Latoszek, Maryn, et al., 2018; de Krom, 1995; Kempster et al., 2009), but many measures such as jitter, harmonics-to-noise ratio, and CPP, which are related to signal periodicity, have been only weakly to moderately correlated with perceived roughness (Barsties v. Latoszek, De Bodt, et al., 2018; Bhuta et al., 2004; de Krom, 1995; Heman-Ackah et al., 2002), similar to our results of CPP and pitch strength. Additionally, our voice samples were selected for being primarily rough with minimal breathiness and strain. The strong relation of perceived roughness with envelope fluctuations may indicate that primary vocal roughness is more strongly related to temporal envelope fluctuations than to signal periodicity. However, signal periodicity and other acoustic factors such as spectral fluctuations can contribute to perceived roughness to some extent. The two most rough samples in our example of model prediction were underestimated by the model equation and may have other acoustic factors affecting their perceived roughness besides envelope fluctuations.

Limitations

This study analyzed perceived roughness in recordings of sustained vowel productions. Perceptual and acoustic evaluations of sustained vowels are routinely performed in voice clinics but may lack ecological validity. Prediction of perceived roughness in connected speech is likely to be more complicated than sustained vowels, and future studies can investigate possible application of temporal fluctuation measures in connected speech. We also purposely chose our samples in a range of *primary* roughness without other voice qualities as much as possible. Natural dysphonic voices often present different voice qualities together, and thus prediction strength of envelope measures in this study may decrease in dysphonic voices covarying with other voice qualities. However, the purpose of choosing primarily rough voice samples was to investigate a direct relationship between perception and temporal fluctuation measures from the auditory model without influence of any other voice qualities. We believe that this purpose was achieved as we obtained a very strong regression model of perceived roughness in this study, and the results provide valuable insight into characteristics of vocal roughness for future studies of voice quality perception and evaluation. Future studies can evaluate the prediction ability of the temporal fluctuation measures to dysphonic voices samples with a variety of other voice qualifies.

Although we have investigated amplitude modulation for predicting perceived roughness, frequency modulation also has been observed to result in perceived roughness in previous studies (Barsties v. Latoszek, De Bodt, et al., 2018; Fastl & Zwicker, 2007). Frequency modulation produces cycle-by-cycle variation in the period (the inverse of frequency) of the stimulus, whereas the amplitude envelope remains unchanged relative to the original carrier amplitude envelope. With frequency modulation, however, if the frequency excursions extend beyond the width of an auditory critical band or auditory filter bandwidth and the rate of those frequency excursions is greater than about 10 Hz, the output of that critical band fluctuates at the rate of frequency modulation. In that case, the frequency modulation is actually perceived as amplitude modulation (Moore & Sek, 1994; Zwicker, 1952). Thus, we suspect that the degree of amplitude modulation estimated in this study may have already reflected some degree of frequency modulation in the voice stimuli. Future studies can further investigate the relationship between amplitude and frequency modulation and their contributions to perceived roughness.

Conclusions

In this study, we illustrated that the output of a bio-inspired model of auditory temporal envelope perception is strongly correlated with the perception of vocal roughness and that the function representing that relationship can be used to predict the perceived vocal roughness of a novel stimulus set as perceived by novel listeners. Perceived roughness was significantly predicted by the degree of envelope modulation, estimated from the auditory model with modulation filter bank, and the output from a modulation filter with center frequency at 64 Hz best predicted perceived roughness. One possible reason for our results may be that our perception is based on a number of transformations of the acoustic signal as it is processed by the auditory brain prior to generating a percept. Future work is needed to combine predictors of different voice qualities into a single comprehensive model.

Data Availability Statement

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

Acknowledgments

This work was supported by NIH R01 DC009029 (D.A.E. and R.S.). The authors would like to thank Mark D. Skowronski for help with programming auditory models and Katherine A. McAllister for help with recruitment and data collection.

References

Anand, S., Kopf, L. M., Shrivastav, R., & Eddins, D. A. (2019). Objective indices of perceived vocal strain. *Journal of Voice, 33*(6), 838–845. https://doi.org/10.1016/j.jvoice.2018.06.005

ANSI. (2010). *Methods for manual pure-tone threshold audiometry*. American National Standards Institute.

Awan, S. N., & Awan, J. A. (2020). A two-stage cepstral analysis procedure for the classification of rough voices. *Journal of Voice, 34*(1), 9–19. https://doi.org/10.1016/j.jvoice.2018.07.003

Awan, S. N., Solomon, N. P., Helou, L. B., & Stojadinovic, A. (2013). Spectral-cepstral estimation of dysphonia severity: External validation. *Annals of Otology, Rhinology & Laryngology, 122*(1), 40–48. https://doi.org/10.1177/000348941312200108

Barsties v. Latoszek, B., De Bodt, M., Gerrits, E., & Maryn, Y. (2018). The exploration of an objective model for roughness with several acoustic markers. *Journal of Voice, 32*(2), 149–161. https://doi.org/10.1016/j.jvoice.2017.04.017

Barsties v. Latoszek, B., Maryn, Y., Gerrits, E., & De Bodt, M. (2018). A meta-analysis: Acoustic measurement of roughness and breathiness. *Journal of Speech, Language, and Hearing Research, 61*(2), 298–323. https://doi.org/10.1044/2017_JSLHR-S-16-0188

Behrman, A. (2005). Common practices of voice therapists in the evaluation of patients. *Journal of Voice, 19*(3), 454–469. https://doi.org/10.1016/j.jvoice.2004.08.004

Bhuta, T., Patrick, L., & Garnett, J. D. (2004). Perceptual evaluation of voice quality and its correlation with acoustic measurements. *Journal of Voice, 18*(3), 299–304. https://doi.org/10.1016/j.jvoice.2003.12.004

Boersma, P., & Weenink, D. (2021). *Praat: Doing phonetics by computer* (Version 6.1.50). http://www.praat.org/

Camacho, A. (2012). *On the use of auditory models' elements to enhance a sawtooth waveform inspired pitch estimator on telephone-quality signals*. 11th International Conference on Information Science, Signal processing and their Applications (ISSPA).

Carding, P. N., Wilson, J. A., MacKenzie, K., & Deary, I. J. (2009). Measuring voice outcomes: State of the science review. *The Journal of Laryngology & Otology, 123*(8), 823–829. https://doi.org/10.1017/S0022215109005398

Dau, T. (2008). Auditory processing models. In D. Havelock, S. Kuwano, & M. Vorländer (Eds.), *Handbook of signal processing in acoustics*. Springer. https://doi.org/10.1007/978-0-387-30441-0_12

Dau, T., Kollmeier, B., & Kohlrausch, A. (1997a). Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers. *The Journal of the Acoustical Society of America, 102*(5), 2892–2905. https://doi.org/10.1121/1.420344

Dau, T., Kollmeier, B., & Kohlrausch, A. (1997b). Modeling auditory processing of amplitude modulation. II. Spectral and temporal integration. *The Journal of the Acoustical Society of America, 102*(5), 2906–2919. https://doi.org/10.1121/1.420345

Dau, T., Puschel, D., & Kohlrausch, A. (1996). A quantitative model of the "effective" signal processing in the auditory system. I. Model structure. *The Journal of the Acoustical Society of America, 99*(6), 3615–3622. https://doi.org/10.1121/1.414959

de Krom, G. (1995). Some spectral correlates of pathological breathy and rough voice quality for different types of vowel fragments. *Journal of Speech and Hearing Research, 38*(4), 794–811. https://doi.org/10.1044/jshr.3804.794

Eddins, D. A., Anand, S., Camacho, A., & Shrivastav, R. (2016). Modeling of breathy voice quality using pitch-strength estimates. *Journal of Voice, 30*(6), 774.e1–774.e7. https://doi.org/10.1016/j.jvoice.2015.11.016

Eddins, D. A., Kopf, L. M., & Shrivastav, R. (2015). The psychophysics of roughness applied to dysphonic voice. *The Journal of the Acoustical Society of America, 138*(6), 3820–3825. https://doi.org/10.1121/1.4937753

Eddins, D. A., & Shrivastav, R. (2013). Psychometric properties associated with perceived vocal roughness using a matching task. *The Journal of the Acoustical Society of America, 134*(4), EL294–EL300. https://doi.org/10.1121/1.4819183

Fastl, H., & Zwicker, E. (2007). *Psychoacoustics: Facts and models* (3rd ed.). Springer.

Heman-Ackah, Y. D., Heuer, R. J., Michael, D. D., Ostrowski, R., Horman, M., Baroody, M. M., Hillenbrand, J., & Sataloff, R. T. (2003). Cepstral peak prominence: A more reliable measure of dysphonia. *Annals of Otology, Rhinology & Laryngology, 112*(4), 324–333. https://doi.org/10.1177/000348940311200406

Heman-Ackah, Y. D., Michael, D. D., & Goding, G. S., Jr. (2002). The relationship between cepstral peak prominence and selected parameters of dysphonia. *Journal of Voice, 16*(1), 20–27. https://doi.org/10.1016/s0892-1997(02)00067-x

Hillenbrand, J., Cleveland, R. A., & Erickson, R. L. (1994). Acoustic correlates of breathy vocal quality. *Journal of Speech and Hearing Research, 37*(4), 769–778. https://doi.org/10.1044/jshr.3704.769

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

Kempster, G. B., Gerratt, B. R., Verdolini Abbott, K., Barkmeier-Kraemer, J., & Hillman, R. E. (2009). Consensus auditory-perceptual evaluation of voice: Development of a standardized clinical protocol. *American Journal of Speech-Language Pathology, 18*(2), 124–132. https://doi.org/10.1044/1058-0360(2008/08-0017)

Kreiman, J., & Gerratt, B. R. (1998). Validity of rating scale measures of voice quality. *The Journal of the Acoustical Society of America, 104*(3), 1598–1608. https://doi.org/10.1121/1.424372

Kreiman, J., & Gerratt, B. R. (2005). Perception of aperiodicity in pathological voice. *The Journal of the Acoustical Society of America, 117*(4), 2201–2211. https://doi.org/10.1121/1.1858351

Langner, G. (1992). Periodicity coding in the auditory system. *Hearing Research, 60*(2), 115–142. https://doi.org/10.1016/0378-5955(92)90015-f

Majdak, P., Hollomey, C., & Baumgartner, R. (2021). *AMT 1.0: The toolbox for reproducible research in auditory modeling* [submitted to Acta Acustica].

Maryn, Y., De Bodt, M., & Roy, N. (2010). The Acoustic Voice Quality Index: Toward improved treatment outcomes assessment in voice disorders. *Journal of Communication Disorders, 43*(3), 161–174. https://doi.org/10.1016/j.jcomdis.2009.12.004

Maryn, Y., & Weenink, D. (2015). Objective dysphonia measures in the program Praat: Smoothed cepstral peak prominence and acoustic voice quality index. *Journal of Voice, 29*(1), 35–43. https://doi.org/10.1016/j.jvoice.2014.06.015

MEEI Voice and Speech Laboratory. (1994). *Disordered Voice Database Model 4337* (Ver. 1.03) [CD-ROM]. Kay Elemetrics Corp.

Mehta, D. D., & Hillman, R. E. (2008). Voice assessment: Updates on perceptual, acoustic, aerodynamic, and endoscopic imaging methods. *Current Opinion in Otolaryngology & Head and Neck Surgery, 16*(3), 211–215. https://doi.org/10.1097/MOO.0b013e3282fe96ce

Moore, B. C. J., & Glasberg, B. R. (1996). A revision of Zwicker's loudness model. *Acustica, 82*(2), 335–345.

Moore, B. C., & Sek, A. (1994). Effects of carrier frequency and background noise on the detection of mixed modulation. *The Journal of the Acoustical Society of America, 96*(2), 741–751. https://doi.org/10.1121/1.410312

Münkner, S. (1993). *Modellentwicklung und Messungen zur Wahrnehmung nichtstationärer akustischer Signale* [Model development and experiments on the perception of nonstationary acoustic signals] [Ph.D. thesis, University of Göttingen].

Patel, R. R., Awan, S. N., Barkmeier-Kraemer, J., Courey, M., Deliyski, D., Eadie, T., Paul, D., Svec, J. G., & Hillman, R. (2018). Recommended protocols for instrumental assessment of voice: American Speech-Language-Hearing Association expert panel to develop a protocol for instrumental assessment of vocal function. *American Journal of Speech-Language Pathology, 27*(3), 887–905. https://doi.org/10.1044/2018_AJSLP-17-0009

Patel, S., Shrivastav, R., & Eddins, D. A. (2010). Perceptual distances of breathy voice quality: A comparison of psychophysical methods. *Journal of Voice, 24*(2), 168–177. https://doi.org/10.1016/j.jvoice.2008.08.002

Patel, S., Shrivastav, R., & Eddins, D. A. (2012). Identifying a comparison for matching rough voice quality. *Journal of Speech, Language, and Hearing Research, 55*(5), 1407–1422. https://doi.org/10.1044/1092-4388(2012/11-0160)

Patterson, R. D., Unoki, M., & Irino, T. (2003). Extending the domain of center frequencies for the compressive gammachirp auditory filter. *The Journal of the Acoustical Society of America, 114*(3), 1529–1542. https://doi.org/10.1121/1.1600720

Shrivastav, R. (2003). The use of an auditory model in predicting perceptual ratings of breathy voice quality. *Journal of Voice, 17*(4), 502–512. https://doi.org/10.1067/s0892-1997(03)00077-8

Shrivastav, R., Eddins, D. A., & Anand, S. (2012). Pitch strength of normal and dysphonic voices. *The Journal of the Acoustical Society of America, 131*(3), 2261–2269. https://doi.org/10.1121/1.3681937

Shrivastav, R., & Sapienza, C. M. (2003). Objective measures of breathy voice quality obtained using an auditory model. *The Journal of the Acoustical Society of America, 114*(4), 2217–2224. https://doi.org/10.1121/1.1605414

Shrivastav, R., Sapienza, C. M., & Nandur, V. (2005). Application of psychometric theory to the measurement of voice quality using rating scales. *Journal of Speech, Language, and Hearing Research, 48*(2), 323–335. https://doi.org/10.1044/1092-4388(2005/022)

Watts, C. R., Awan, S. N., & Maryn, Y. (2017). A comparison of cepstral peak prominence measures from two acoustic analysis programs. *Journal of Voice, 31*(3), 387.e1–387.e10. https://doi.org/10.1016/j.jvoice.2016.09.012

Zwicker, E. (1952). Die Grenzen der Hörbarkeit der Amplitudenmodulation und der Frequenzmodulation eines Tones [The limits of audibility of amplitude modulation and frequency modulation of a pure tone]. *Acustica, 2,* 125–133.