JSLHR

Research Article

Perceptual and Acoustic Assessment of Strain Using Synthetically Modified Voice Samples

Yeonggwang Park,[a] Manuel Díaz Cádiz,[a] Kathleen F. Nagle,[b] and Cara E. Stepp[a,c,d]

Purpose: Assessment of strained voice quality is difficult due to the weak reliability of auditory-perceptual evaluation and lack of strong acoustic correlates. This study evaluated the contributions of relative fundamental frequency (RFF) and mid-to-high frequency noise to the perception of strain.
Method: Stimuli were created using recordings of speakers producing /ifi/ with a comfortable voice and with maximum vocal effort. RFF values of the comfortable voice samples were synthetically lowered, and RFF values of the maximum vocal effort samples were synthetically raised. Mid-to-high frequency noise was added to the samples. Twenty listeners rated strain in a visual sort-and-rate task. The effects of RFF modification and added noise on strain were assessed using an analysis of variance; intra- and interrater reliability were compared with and without noise.
Results: Lowering RFF in the comfortable voice samples increased their perceived strain, whereas raising RFF in the maximum vocal effort samples decreased their strain. Adding noise increased strain and decreased intra- and interrater reliability relative to samples without added noise.
Conclusions: Both RFF and mid-to-high frequency noise contribute to the perception of strain. The presence of dysphonia may decrease the reliability of auditory-perceptual evaluation of strain, which supports the need for complementary objective assessments.
Supplemental Material: https://doi.org/10.23641/asha.13172252

The lifetime prevalence of voice disorders in the United States is 30%, with an incidence of 7% (Roy et al., 2005). Voice disorders disrupt an individual's quality of life, affecting both economic and social activities (Smith et al., 1996). One of the most common features of voice disorders is vocal hyperfunction (Stemple et al., 2014), which comprises approximately 65% of all cases in voice clinics in the United States (Brodnitz, 1966; Ramig & Verdolini, 1998). Vocal hyperfunction involves excessive and/or imbalanced laryngeal and paralaryngeal muscular forces and is often associated with phonotrauma, which can result in organic changes to the vocal folds (Hillman et al., 1989). Vocal hyperfunction is also present without phonotrauma; this type of vocal hyperfunction is usually referred to as muscle tension dysphonia, which is estimated to comprise 10%–40% of vocal disorders diagnosed clinically (Roy, 2003).

The current clinical assessment of vocal hyperfunction is primarily based on auditory-perceptual evaluation (Roy et al., 2013). Auditory-perceptual evaluation of voice quality is routinely used in clinical assessment (Carding et al., 2009; De Bodt et al., 1996) due to its convenience and efficiency (Kent, 1996). It is currently considered the gold standard for evaluating the severity of voice disorders and the outcome of voice therapy (Oates, 2009); thus, auditory-perceptual evaluation is essential to clinical management of voice disorders. A major auditory-perceptual quality associated with vocal hyperfunction is strain. It is defined as the "perception of excessive vocal effort (hyperfunction)" in a standard clinical tool, the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V; Kempster et al., 2009). Because strain is a major perceptual attribute of vocal hyperfunction (Kempster et al., 2009; Morrison, 1997), evaluating strain is important to guide the treatment of individuals with voice disorders related to vocal hyperfunction.

[a]Department of Speech, Language and Hearing Sciences, Boston University, MA
[b]Department of Speech-Language Pathology, Seton Hall University, South Orange, NJ
[c]Department of Biomedical Engineering, Boston University, MA
[d]Department of Otolaryngology – Head and Neck Surgery, Boston University School of Medicine, MA
Correspondence to Yeonggwang Park: yppark@bu.edu

Editor-in-Chief: Bharath Chandrasekaran
Editor: Kate Bunton
Received May 28, 2020
Revision received July 23, 2020
Accepted August 17, 2020
https://doi.org/10.1044/2020_JSLHR-20-00294

Disclosure: *Cara E. Stepp has received consulting fees from Altec, Inc./Delsys, Inc., companies focused on developing and commercializing technologies related to human movement. Stepp's interests were reviewed and are managed by Boston University in accordance with their conflict of interest policies. All other authors have declared that no competing interests existed at the time of publication.*

Despite the reliance on it in voice clinics, auditory-perceptual evaluation has low reliability. Because of the inherently subjective nature of perceptual evaluation, highly experienced listeners frequently disagree with one another when rating voice quality (Kreiman et al., 1993). This disagreement seems to affect the evaluation of strain more than other voice qualities such as breathiness and roughness, resulting in particularly poor intra- and interrater reliability (Webb et al., 2004; Zraick et al., 2011). Accurate evaluation of different disordered voice qualities can assist clinicians in choosing the most appropriate therapy technique targeted to an individual (Stemple & Hapner, 2019), suggesting that better methods of evaluating strain have the potential to improve clinical outcomes.

Instrumental measures are often used to supplement auditory-perceptual ratings, but there is no strong acoustic correlate of strain yet available. The smoothed cepstral peak prominence (CPPS) is a cepstral peak amplitude normalized over the entire background signal amplitude calculated from the smoothed cepstrum. CPPS has been shown to correlate strongly with auditory-perceptual ratings of overall severity of dysphonia and breathiness (Awan & Roy, 2006; Heman-Ackah et al., 2003; Hillenbrand et al., 1994). Cepstral measures related to CPPS have also shown potential for assessing roughness (Awan & Awan, 2020). However, cepstral measures have shown mixed results for strain (Anand et al., 2019; Lowell et al., 2012; McKenna & Stepp, 2018; Van Stan et al., 2020), and strain has not been strongly correlated with other conventional acoustic measures such as frequency and amplitude perturbation measures (Bhuta et al., 2004).

One potential reason for difficulty with perceptual and acoustic evaluation of strain may be its multidimensionality. It has been shown that breathiness and roughness often accompany strain (Lowell et al., 2012) and that ratings of strain are likely influenced by other co-occurring voice qualities (Kent, 1996). Thus, to improve the auditory-perceptual and acoustic evaluation of strain, its acoustic factors must be revealed. Three acoustic characteristics related to strain have been suggested previously: increased spectral energy at higher harmonic frequencies (Anand et al., 2019; Bergan et al., 2004; Klatt & Klatt, 1990; Stevens, 1977; Sundberg & Gauffin, 1978), decreased relative fundamental frequency (RFF; Stepp et al., 2010, 2011), and increased mid-to-high frequency noise (Hirano, 1981; Klatt & Klatt, 1990; Lowell et al., 2012).

Increased Spectral Energy at Higher Harmonic Frequencies

Increased spectral energy at higher harmonic frequencies has been associated with strain (Anand et al., 2019). Increased energy at higher harmonic frequencies has also been associated with a pressed voice quality, which results from phonation with excessively adducted vocal folds, suggesting a similarity between strained and pressed voice qualities (Kreiman et al., 2012). Increases in energy at higher harmonic frequencies in synthesized voice samples increased listeners' perceptions of pressed voice quality (Bergan et al., 2004). Thus, the relationship between increased spectral intensity at higher harmonic frequencies and strain is well supported, both theoretically and empirically.

Decreased RFF

RFF has been proposed as an acoustic feature reflecting increased laryngeal tension and strain (Stepp et al., 2010, 2011). RFF quantifies the short-term variation of fundamental frequency (f_o) in sonorant-voiceless consonant-sonorant productions. It is defined as the instantaneous f_os of the 10 voicing offset and onset cycles before and after a voiceless consonant, normalized by the f_os of the cycles furthest from the consonant. Compared to the RFF of individuals with healthy voices, RFF values are lower in individuals thought to have increased laryngeal tension, including those with vocal hyperfunction (Heller Murray et al., 2017; Roy et al., 2016; Stepp et al., 2010, 2012), Parkinson disease (Bowen et al., 2013; Goberman & Blomgren, 2008; Stepp, 2013), and adductory laryngeal dystonia (Eadie & Stepp, 2013). Stepp et al. (2010) hypothesized that increased baseline laryngeal tension would decrease the extent of the f_o changes before and after intervocalic voiceless consonant production.

The relationship between RFF and strain has also been evaluated and was found to be moderate in previous auditory-perceptual studies (Eadie & Stepp, 2013; Lien et al., 2015; McKenna & Stepp, 2018; Stepp et al., 2012). However, it is not clear whether listeners responded specifically to the changes in RFF or other acoustic features that may change in concert with RFF. In addition, changes in RFF values over a period of high voice use did not result in changes in strain perceived by listeners in one study (Heller Murray et al., 2016). This suggests that RFF may reflect underlying laryngeal tension that may not necessarily be perceived by listeners. Thus, it is not yet clear whether RFF is directly perceived by listeners as changes in strain or whether it covaries with other acoustic features that are perceived by listeners.

Increased Mid-to-High Frequency Noise

Strain has also been described as containing increased mid-to-high frequency noise (Hirano, 1981) and as being often accompanied by perceived breathiness (Lowell et al., 2012). Breathiness is known to be associated with increased aspiration noise in the mid-to-high frequency range near the third formant (Klatt & Klatt, 1990). However, if aspiration noise interferes with higher harmonic frequencies in a similar frequency range, which may also contribute to strained voice quality, it is unclear how aspiration noise actually affects strain. Kreiman and Gerratt (2012) observed that increases in noise decreased listeners' acuity to changes in the harmonic structure of the voice source. Thus, the presence of noise may also decrease listeners' acuity to the percept caused by RFF, which is also dependent on the ability of the peripheral auditory system to resolve the harmonic structures of the voice source. In summary,

aspiration noise may interfere with other acoustic characteristics of strain, and it is thus unclear how aspiration noise may contribute to the perception of strain. In order to study the effect of aspiration noise, mid-to-high frequency noise can be synthetically added to voice samples.

Purpose

In this study, we aimed to understand the contributions of the two acoustic characteristics, RFF and mid-to-high frequency noise, to the auditory-perceptual measure of strain. We used synthesis techniques to precisely control these acoustic features and evaluate their direct associations with strain. The previously observed correlations with natural voice samples cannot fully elucidate the relationships between these acoustic measures and auditory-perception, as other acoustic parameters may also differ across voice samples. By modifying only the acoustic parameters of interest, we aimed to delineate more directly the roles of RFF and mid-to-high frequency noise on the auditory-perceptual evaluation of strain. We did not examine increased spectral energy at higher harmonic frequencies in this study because it has already been examined with synthesized samples and showed a statistically significant association with strain (Bergan et al., 2004).

We hypothesized that synthetically lowering RFF would result in an increase in the perception of strain and that synthetically raising RFF would result in a decrease in the perception of strain. We also hypothesized that adding mid-to-high frequency noise to speech samples would increase the perception of strain. We further hypothesized that adding noise would result in decreases in both intra- and interrater reliability of strain ratings, since noise may interfere with other acoustic characteristics of strain.

Method

Original Voice Recordings

Voice samples of eight individuals (four women and four men; M_{age} = 32.6 years, range: 18–67 years) with healthy voices were selected from a database of participant recordings of RFF stimuli, /ifi/. These recordings were collected from speakers who were asked to increase their vocal effort to mild, moderate, and maximum levels. The speakers were given the instruction, "produce your voice as if you are trying to push out the air without increasing the loudness," and the experimenter provided demonstrations of different vocal effort levels. Recordings of comfortable voice and maximum vocal effort conditions were used because they were expected to show the largest differences in strain and RFF values among all combinations of the recordings. Three /ifi/ productions were selected from each effort condition for each participant. Recordings were selected for inclusion based on three criteria to best support the study hypotheses: (a) increases in self-modulated vocal effort accompanied with decreases in RFF, (b) increases in self-modulated vocal effort accompanied with increases in listener-perceived strain, and (c) minimal listener-perceived breathiness regardless of vocal effort level. These criteria are further explained below.

Increases in Self-Modulated Vocal Effort Accompanied With Decreases in RFF

RFF is generally expected to decrease as vocal effort increases, although between-speaker variability has been reported (Lien et al., 2015; McKenna & Stepp, 2018; Stepp et al., 2012). We purposefully chose voice samples that showed decreased RFF along with increased vocal effort to evaluate the contribution of RFF to strain. To achieve this aim, we planned to synthetically lower RFF values of comfortable voice samples to match the RFF values of maximum effort samples from the same speakers and to examine whether the RFF modifications increased the strain. We also planned to synthetically raise RFF values in maximum effort samples to match the RFF values of comfortable voice samples from the same speakers and to examine whether the RFF modifications decreased the strain. RFF was manually estimated with Praat acoustic analysis software (Version 6.0.48; Boersma & Weenink, 2019). Ten voiced cycles prior to and after the voiceless consonant were identified using Praat's autocorrelation algorithm for pitch tracking, and the period and the instantaneous f_o for each cycle were calculated. RFF for each cycle was calculated in semitones (ST) from the equation: $ST = 39.86 \times \log_{10}(f_o/\text{reference } f_o)$, in which the reference f_o (fref) was offset cycle 1 for offset cycles and onset cycle 10 for onset cycles. Mean RFF was higher in comfortable voice samples than in maximum vocal effort samples.

Increases in Self-Modulated Vocal Effort Accompanied With Increases in Listener-Perceived Strain

We also chose recordings in which the strain increased as the self-modulated vocal effort level increased. Strain of the recordings was evaluated by a voice-experienced speech-language pathologist. Since speakers could increase their vocal effort without actual perceptible increases in strain by listeners, this criterion ensured that the comfortable voice and maximum vocal effort samples had actual differences in strain. The speech-language pathologist rated the strain of each recording on a 100-mm visual analog scale from the CAPE-V form (Kempster et al., 2009). Mean strain was higher in maximum vocal effort samples than in comfortable voice samples.

Minimal Listener-Perceived Breathiness Regardless of Vocal Effort Level

Speakers with minimal listener-perceived breathiness were favored because of the study aim to evaluate the effect of mid-to-high frequency noise on perceiving strain and RFF. Since we planned to compare samples with and without added noise, recordings with minimally breathy voices would be ideal to precisely control for noise. Minimal breathiness was determined from both the auditory-perceptual evaluation by a voice-experienced speech-language pathologist and CPPS values. The speech-language pathologist rated

the breathiness of each recording on a 100-mm visual analog scale from the CAPE-V form (Kempster et al., 2009) after listening to three /ifi/ samples per speaker. CPPS represents the strength of the cepstral peak compared to the cepstral background noise in acoustic signals, which reflects the degree of the periodicity of the signal. CPPS has shown a strong negative correlation with perceived breathiness (Hillenbrand et al., 1994). CPPS was obtained from the /i/ portions of /ifi/ recordings with the commands and parameters described in Watts et al. (2017). Mean breathiness and CPPS values were similar to those of 20 young female adults with healthy voices in our previous study (Park et al., 2019) and did not differ between the two vocal effort conditions.

Stimuli Synthesis

In order to synthetically modify the selected recordings, we used the Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum (STRAIGHT) algorithm. The STRAIGHT is based on a sophisticated channel VOCODER system, which separates the spectral and source information (Kawahara, 2006). The algorithm incorporates speech analysis, modification, and synthesis. During the analysis, the algorithm extracts information about the spectral envelope, the instantaneous f_o contour, and the aperiodic component of a sound sample. These three components can be modified separately and then synthesized back together to generate a modified voice sample. This method was developed to provide flexible modification of the three components as naturally as possible and was adapted to MATLAB (Version. R2018a, MathWorks). The 16 original recordings were analyzed, and the STRAIGHT components were saved for sample modifications. Instead of including the original recordings as a part of the perceptual stimuli, we used STRAIGHT-synthesized versions of the original recordings. The STRAIGHT components from the original recordings were resynthesized back without any modification to be included in the experiment as *unmodified* samples. This ensured that possible perceptual differences between the original samples and RFF modified samples would not be due to being synthesized from the STRAIGHT, although the original recordings and the STRAIGHT-synthesized versions are known to be perceptually identical (Kawahara, 2006). Each sample consisted of three /ifi/s in the same vocal effort condition from the same participant with 300-ms periods between each /ifi/.

Modifying RFF

RFF of the unmodified samples was modified in order to test the hypothesis that the modification of RFF alone would alter the strain. Modifying RFF of the samples and comparing the RFF-modified and unmodified samples allowed precise evaluation of RFF in relation to strain, since RFF was the only acoustic feature that was different between them. RFF of the comfortable voice sample from each participant was lowered to the RFF values of the same participant's maximum vocal effort sample for all 20 RFF cycles. RFF of the maximum vocal effort sample from each participant was raised to the RFF values of the same participant's comfortable vocal effort sample for all 20 RFF cycles. In order to modify RFF, we modified the f_o contours of the samples, as estimated by the STRAIGHT. Figure 1 illustrates the procedure for modifying f_o contours. The modified f_o contours and other STRAIGHT components of the same sample were combined together to synthesize an RFF-modified sample. We performed the RFF modification on all unmodified samples and confirmed that the modified RFF values matched the goal RFF values very closely. The mean difference in RFF between comfortable voice samples with RFF modification and maximum vocal effort samples without RFF modification was 0.14 ST; the mean difference in RFF between maximum vocal effort samples with RFF modification and comfortable voice samples without RFF modification was 0.01 ST. Because the spectral envelopes of the samples were not modified during the process, the formant values of RFF-modified samples were not altered.

Adding Mid-to-High Frequency Noise

Versions of samples with and without RFF modification with added mid-to-high frequency noise were created via the "breathiness" function in the Praat Vocal Toolkit (Corretge, 2019). This function uses linear predictive coding to estimate the spectral envelope of the original signal. The function then creates a "whispered" version of the signal by applying the estimated spectral envelope to white noise, decreasing the spectral energy of the low frequencies under 250 Hz, and increasing the spectral energy of the mid-to-high frequencies centered around 2000 Hz. The whispered version of the signal is then added to the original signal in a quantity determined by the user's input. In order to synthesize the samples to be breathy while retaining a natural quality, we decreased the harmonics-to-noise ratio (HNR) values of the samples by −7 dB using the breathiness function. The mean HNR of the samples without added noise was 19.3 dB, and the mean HNR of the samples with added noise was 12.3 dB. A speech-language pathologist evaluated the breathiness of all samples; the breathiness ratings increased by an average of 30.6 mm on the 100-mm scale in the samples with added noise.

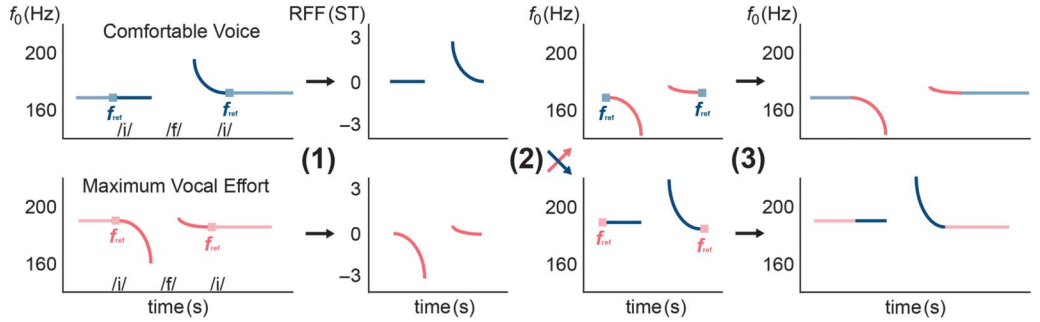
Total Number of Stimuli

The number of the STRAIGHT-synthesized, unmodified samples was 16 (8 participants × 2 vocal effort conditions). RFF was modified in all unmodified samples, and noise was added in both RFF-modified and unmodified samples, resulting in a total of 64 samples for the perceptual tasks (16 unmodified samples + 16 RFF-modified samples + 32 samples with added noise [unmodified and RFF-modified]).

Listeners

Twenty healthy participants (10 women and 10 men; M_{age} = 22.0 years; range: 18–34 years) were recruited as

Figure 1. The procedures for modifying fundamental frequency (f_o) contours of the comfortable and maximum vocal effort samples to exchange their RFF values. The upper panels present schematic plots of f_o (in Hertz [Hz]) and RFF (in semitones [ST]) as a function of time (in seconds [s]). Blue represents a comfortable voice sample and pink represents a maximum vocal effort sample. RFF portions of the f_o contour contain bolder colors than non-RFF portions (abbreviation: f_{ref} = f_o of the reference cycle; RFF = relative fundamental frequency). The following steps outline the details of the procedures: (1) RFF portions were selected from the f_o contours of the comfortable voice and maximum vocal effort samples. The selected f_o contour of each sample was normalized by the f_{ref} values of each sample (RFF [ST] = $39.84 \times \log_{10}[f_o/f_{ref}]$). (2) The RFF contours of the comfortable and maximum vocal effort voice samples were exchanged and transformed back to f_o contours that fit their respective counterpart's f_{ref} values (f_o [Hz] = $f_{ref} \times 10\text{^}[RFF/39.84]$). (3) The converted f_o contours of the comfortable and maximum vocal effort samples replaced the RFF portions of their respective counterpart's f_o contours. During this process, durations of the converted f_o contours were adjusted, so that the original durations of RFF portions remained the same after RFF modification.



listeners from college job posting sites and paper flyers and were paid for their participation. The number of participants was determined by evaluation of the average absolute deviations of strain from mean strain ratings obtained using a visual sort-and-rate (VSR) task from 20 listeners in a previous study (McKenna & Stepp, 2018). Mean strain rated by 18 listeners differed from mean strain rated by 20 participants by 1 mm on a 100-mm scale. We recruited 20 participants to attain similar precision. Participants reported no prior history of speech, language, and hearing disorders or previous participation in any auditory-perceptual study. Participants all scored within normal ranges for the Voice-Related Quality of Life (Hogikyan & Sethuraman, 1999). All but one participant passed a hearing screening with 25 dB HL pure tones at 125, 250, 500, 1000, 2000, 4000, and 8000 Hz (American Speech-Language-Hearing Association, 2005) in a sound-treated room (one participant passed at 30 dB HL at 4000 Hz in his left ear). Inexperienced listeners were recruited, as previous studies did not find differences in interrater reliability values between expert and inexperienced listeners (Eadie et al., 2010).

Perceptual Tasks

VSR Task for Strain

The participants completed VSR training and experimental tasks in a sound-treated room. The VSR task was chosen because of its higher reliability compared to other auditory-perceptual tasks (Granqvist, 2003). Participants were provided the CAPE-V definition of strain, "perception of excessive vocal effort (hyperfunction)" (Kempster et al., 2009), and the definition of vocal effort, "perceived exertion in producing voice" (Verdolini et al., 1994).

First, they were trained to use the VSR module on a desktop computer and familiarized themselves with a wide range of strain. The training module included eight voice samples, each containing three /ifi/s with 300-ms breaks, similar to experimental stimuli. The eight training samples were chosen from the same database of original recordings as the experimental stimuli. The eight training samples were selected to contain a wide range of strain based on strain ratings from three voice-experienced speech-language pathologists. We averaged the strain ratings from the three raters in order to improve the reliability of the training set. Strain ratings of the training set ranged from 2.2 to 83.4 mm, spread evenly throughout the 100-mm range. At the start of the training, icons for the samples were located horizontally at the middle of the vertical axis, which ranged from 0 mm (no strain) to 100 mm (the most strain). When the participants clicked each icon on the screen, the corresponding sample was presented at 75 dB SPL through a pair of Sennheiser HD-290 headphones. Participants were allowed to listen to the samples as many times as they wished. They were asked to first listen to each stimulus and rate the strain by moving icons vertically on the strain scale. After finishing the initial listening and rating the samples of the training set, participants were asked to relisten to each sample and adjust their ratings by comparing the samples that were located near each other vertically. When the participants finished rating the training set, they were given the experts' scores of the training samples as feedback, so that they could familiarize themselves with the experts' ratings on these training samples. This familiarization with the experts' ratings was aimed at improving the interrater reliability of the task, as poor interrater reliability of strain ratings has been previously reported (Webb et al., 2004; Zraick et al., 2011).

After the training module, listeners were asked to complete the experimental VSR module, which contained the same screen setup as the training module. A total of 80 stimuli, the 64 stimuli and 16 randomly chosen stimuli

from the 64 stimuli for intrarater reliability, were divided into 10 sets of eight stimuli. Each set was designed to contain only one stimulus from each speaker so that the samples from the same speaker would not be compared to each other within a set. Each set was also designed to contain one stimulus from each modification type (e.g., RFF-modified comfortable voice samples with added noise, RFF-unmodified maximum vocal effort samples with added noise) so that every set would contain samples with a wide range of strain. Each set also contained at least one repeated stimulus from the other sets for intrarater reliability assessment. The 10 sets of stimuli were constructed specifically for each listener to reduce the effects of stimuli order and set composition. The experimental VSR task took approximately 15 min to complete.

Same or Different Task

Participants completed an AX (same or different) task in a sound-treated room to evaluate whether listeners could differentiate between the samples that differed only in their RFF. Each /ifi/ in the RFF-unmodified samples was paired with its own RFF-modified version with a 300-ms interstimulus interval. A total of 96 pairs of RFF-modified and unmodified samples were possible from the three /ifi/s in our 32 RFF-modified and 32 RFF-unmodified samples. Within each pair, the order of RFF-modified and unmodified samples were randomly determined. In order to balance the number of same and different trials in the task, 96 stimuli pairs with the same /ifi/s were randomly chosen from the stimuli set and included in the task. After hearing each pair of stimuli, listeners judged whether the two stimuli were same or different in a forced-choice paradigm. They were asked to listen very carefully and were informed that the difference in the two samples could be very small, but they were not given any information about the basis of any differences. In total, 192 trials were performed by each listener, taking approximately 20 min to complete.

Data Analysis

Strain ratings for each stimulus obtained from the VSR tasks were averaged across the listening participants. The number of correct "different" responses of each participant was obtained from the AX task, and the correct response rate was calculated for each of the four stimulus conditions: comfortable voice samples with and without noise and maximum vocal effort samples with and without noise. The number of wrong "different" response was also obtained, and the false-alarm rate was calculated for each stimulus condition. The sensitivity index, *d'*, was calculated from the equation below presented in Macmillan and Creelman (2004) for each stimulus condition of each listener: $d' = z(\text{correct response rate}) - z(\text{false-alarm rate})$, where z is the inverse of the normal distribution. When either rate was 0 (which inhibits the calculation of *d'*), we used $1/(2 \times$ the number of trials in a stimulus condition [24]) instead. A high, positive *d'* value would indicate high discriminability between RFF-modified and unmodified samples, whereas a zero *d'* value would indicate chance level performance (Macmillan & Creelman, 2004).

Statistical Analysis

Statistical analysis was performed in SPSS (Version. 24.0, IBM Corp.). A three-way repeated-measures analysis of variance (ANOVA) on the mean strain of the stimuli was performed with the factors: vocal effort level (comfortable voice or maximum vocal effort), RFF modification (unmodified or modified), noise (no or added noise), and the interactions between the factors. We hypothesized that the interaction between vocal effort level and RFF modification would be statistically significant, which would support the contribution of RFF to strain. We did not hypothesize a main effect of RFF modification on strain because the direction of RFF modification depended on vocal effort level: The comfortable voice samples would have increased strain due to their lowered RFF, whereas the maximum vocal effort samples would have decreased strain due to their raised RFF. We also hypothesized that either noise or the interaction between noise and vocal effort level would be statistically significant because mid-to-high frequency noise was expected to increase strain. Finally, we hypothesized that there would be a statistically significant interaction between noise, vocal effort level, and RFF modification because noise may affect the listeners' acuity of the percept caused by RFF. We also performed a two-way repeated-measures ANOVA on *d'* values from the AX task with vocal effort level and noise as factors to evaluate how adding the noise would affect the ability of listeners to notice differences in RFF. Effect sizes were calculated as a partial eta squared (η_p^2), and post hoc tests were performed when statistically significant interactions were observed.

Intrarater reliability of the ratings of strain was assessed using Pearson correlations from 16 repeated samples. Strain showed intrarater reliability (Pearson *r*) above .7 in 18 of 20 listeners (*Mdn* = .85, range: .42–.99). Interrater reliability was represented as intraclass correlation coefficient (ICC; two-way mixed effects, consistency, single measure) calculated with the ratings of all 64 stimuli from all listeners. ICC below 0.5 has been considered as poor reliability, .5–.75 as moderate reliability, .75–.9 as good reliability, and above .9 as excellent reliability (Portney & Watkins, 2000). Ratings of strain showed moderate interrater reliability (ICC = .61, 95% CI [.53, .70]).

We additionally evaluated intrarater and interrater reliability separately for samples with and without noise to examine our hypothesis that mid-to-high frequency noise may decrease the reliability of strain rating. Among the 16 repeated samples, half of them were samples with added noise and the other half was without noise. In order to evaluate our hypothesis that mid-to-high frequency noise would decrease intrarater reliability of strain, mean absolute differences in strain between the actual and repeated samples were calculated separately for samples with and without noise from each listener. An independent *t* test was performed on the mean absolute differences in strain

to evaluate whether the samples with noise resulted in a larger mean absolute difference than samples without noise, which would suggest lower intrarater reliability. To evaluate our hypothesis that noise would decrease interrater reliability, we calculated mean absolute deviation by obtaining absolute deviations of an individual listener's strain rating of a sample from the sample's mean strain rating by the 20 listeners and averaging absolute deviations within each sample. A paired *t* test was performed between the mean absolute deviations of the samples with and without added noise to examine if adding noise increased mean absolute deviation, which would suggest decreased interrater reliability. A predetermined level of statistical significance ($\alpha = .05$) was used for all statistical tests.

Results

RFF Modification

The three-way ANOVA on mean strain showed a statistically significant effect of the interaction between vocal effort level and RFF modification with a large effect size ($p = .003$, $\eta_p^2 = .74$). This interaction indicates that RFF modification changed the perceived strain of the samples but that the effect differed based on the vocal effort level. The post hoc paired *t* test between the comfortable voice samples with and without RFF modification revealed that synthetically lowering RFF values in the comfortable voice samples resulted in increases in strain ($t = -5.4$, $p < .001$; see Figure 2), as hypothesized. The post hoc paired *t* test between the maximum vocal effort samples with and without RFF modification revealed that synthetically raising RFF values in the maximum vocal effort samples resulted in decreases in strain ($t = 3.5$, $p = .003$; see Figure 2). The mean *d'*, which represents the listeners' performance discriminating between RFF-modified and unmodified samples on the AX task, ranged from 0.12 to 0.40 ($M = 0.29$, 95% CI [0.17, 0.40]) in the four stimulus conditions, all above 0 in the scale, in which 0 indicates chance-level performance (see Figure 3).

Mid-to-High Frequency Noise

The effect of mid-to-high frequency noise on strain was statistically significant and showed a large effect size ($p < .001$, $\eta_p^2 = .97$). The samples with added noise had increased strain (see Figure 2). There was no statistically significant effect on the interaction between noise, vocal effort, and RFF ($p = .83$). The addition of noise was not a statistically significant factor in the one-way ANOVA on mean *d'* ($p = .18$; see Figure 3).

The mean absolute difference in strain between the samples and their repetitions was statistically greater ($t = 2.45$, $p = .01$) in the samples with noise ($M = 12.2$, 95% CI [11.4, 13.1]) relative to the samples without noise ($M = 8.5$, 95% CI [7.9, 9.0]), suggesting lower intrarater reliability of strain ratings in the samples with noise than without noise. The mean absolute deviation was also statistically greater ($t = 2.20$, $p = .035$, mean difference = 1.4, 95% CI [0.1, 2.7]) in the samples with noise than samples without noise, suggesting lower interrater reliability of strain ratings in the samples with noise than without noise.

Discussion

In this study, we performed auditory-perceptual experiments with synthetically modified voice samples to evaluate direct, causal contributions of RFF and mid-to-high frequency noise to the perception of strain. We hypothesized that synthetically lowering RFF in voice samples would increase strain, whereas raising RFF in voice samples

Figure 2. Mean strain ratings of comfortable and maximum effort samples as a function of modification condition. Error bars indicate 95% confidence intervals, and bolded brackets and asterisks indicate statistically significant differences between RFF-modified and unmodified samples within comfortable voice ($p < .001$) and maximum vocal effort ($p = .003$) conditions. RFF = relative fundamental frequency; N = noise; - = unmodified; ↑ = increase; ↓ = decrease.

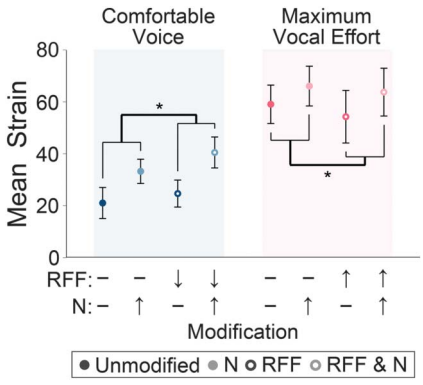
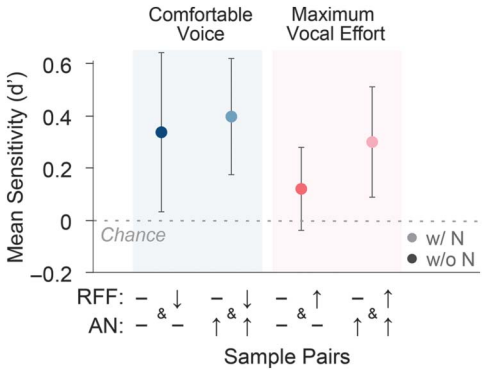


Figure 3. Mean sensitivity of discriminating RFF-modified and unmodified samples from the AX task as a function of paired condition in comfortable voice and maximum vocal effort samples. The addition of noise was not a statistically significant factor on mean *d'* ($p > .05$). A dotted line indicates chance-level discrimination. Error bars indicate 95% confidence intervals. RFF = relative fundamental frequency; w/ N = with noise; w/o N = without noise; - = unmodified; ↑ = increase; ↓ = decrease.



would decrease strain. We also hypothesized that adding mid-to-high frequency noise in voice samples would both increase strain and decrease intra- and interrater reliability of strain ratings.

RFF

The statistically significant interaction between vocal effort level and RFF modification supports the role of RFF as an acoustic contributor to strain. The mean *d'* value greater than 0 from the AX tasks also supports that differences in RFF can be noticed by listeners, although with difficulty (low *d'* values). Our finding is consistent with previously observed correlations between RFF and strain in speakers with healthy voices who modulated their vocal effort (Lien et al., 2015; McKenna & Stepp, 2018), speakers with healthy voices and vocal hyperfunction (Stepp et al., 2012), and speakers with laryngeal dystonia (Eadie & Stepp, 2013). Our findings further support the contribution of RFF to strain by showing that strain changed when only RFF was modified in the acoustic samples while other acoustic features remained constant.

A potential reason for the relationship between RFF and perceived strain may be the high prevalence of decreased RFF in individuals with increased laryngeal tension and vocal effort in their voice production. In our study, decreasing RFF values in the comfortable voice samples resulted in increases in strain. This decreased RFF pattern has been observed in individuals with increased laryngeal tension and vocal effort during their voice production in previous studies (Eadie & Stepp, 2013; Heller Murray et al., 2017; Lien et al., 2015; Roy et al., 2016; Stepp, 2013; Stepp et al., 2010, 2011). Specifically, these individuals showed decreasing offset RFF and slightly increased and then decreasing onset RFF, whereas individuals speaking with typical voices showed stable or slightly decreasing offset RFF and substantially increased and then decreasing onset RFF. Stepp et al. (2011) and Roy et al. (2016) also observed that successful voice therapy sessions normalized this decreased RFF pattern (although Roy et al., 2016, only observed this finding in onset RFF). Individuals with increased laryngeal tension and vocal effort are known to have strained voice quality (Kempster et al., 2009; Roy, 2008), which is suggested to be multidimensional (Kent, 1996; Lowell et al., 2012). Thus, we may frequently encounter the decreased RFF pattern concurrently present with other acoustic features of strain in individuals with increased laryngeal tension and may associate it with increased strain.

Although the contribution of RFF to perceived strain is supported in this study, RFF is probably a small factor in the overall construct of strain due to its short duration and linguistic constraints. Although we observed a statistically significant and large effect of the interaction between vocal effort level and RFF modification, the average change in strain due to RFF modification was small, less than 10 mm on the 100-mm scale. These small changes in strain may have been due to the fact that the RFF measure only spans a small proportion of each utterance, whereas other acoustic features such as increased spectral energies at higher harmonic frequencies and mid-to-high frequency noise can span entire utterances. The short duration of RFF cycles may also explain why it was challenging for the listeners to consistently differentiate between the samples with and without RFF modification in the AX task. The duration of 20 RFF cycles can be estimated from our speakers' average f_os producing /ifi/, which ranged from 105 to 254 Hz. Based on that f_o range, the duration of 20 RFF cycles is estimated to range only from 79 to 190 ms ($1/f_o \times 20$ cycles), whereas the duration of the entire utterance ranged from approximately 400 to 1,000 ms. The proportion of RFF in a sound sample of typical running speech will further decrease, as these are likely not to contain many vowel-consonant-vowel contexts with voiceless consonants. Thus, we hypothesize that, in running speech stimuli, the contribution of RFF to strain also would be even smaller.

Mid-to-High Frequency Noise

The results of the VSR task for rating strain showed that mid-to-high frequency noise is also a statistically significant contributor to strain with a large effect. The effect of noise was also stronger than of RFF, probably because it was present in much longer durations of the samples than RFF. Our finding is consistent with previous findings that showed that increases in breathiness or aspiration noise were coincident with increased strain (Hirano, 1981; Lowell et al., 2012). Listeners may associate increased mid-to-high frequency noise with strain because of the high prevalence of aspiration noise in individuals with strained voices (e.g., individuals with glottal insufficiency, vocal nodules, and paralysis) who need to increase their vocal effort in order to phonate. Aspiration noise also may be perceived as increased respiratory effort, which usually accompanies an increased airflow rate (Zhang, 2015).

There was no statistically significant interaction between noise and RFF. We had predicted that noise would affect the listeners' acuity to the percept caused by RFF. There was also no statistically significant effect of noise on the *d'* in the AX task, which suggests that mid-to-high frequency noise may not affect the ability to notice differences in RFF. However, we may not have observed noise reducing *d'* in the AX task because of the inherent difficulty of the task: The task resulted in overall low values of *d'*, suggesting a floor effect.

Our findings also suggest that mid-to-high frequency noise may decrease both intra- and interrater reliability of strain ratings. These findings are likely not due to noise interfering with the perception of RFF, since the effect of noise on discriminability between samples with and without RFF modification was not statistically significant. Instead, mid-to-high frequency noise is likely to interfere with higher harmonic frequencies, located in a similar frequency range. This speculation is consistent with previous findings from Kreiman and Gerratt (2012) that showed that increased noise in samples reduced sensitivity to harmonic frequencies. Thus, listeners may have perceived

different amounts of energy at higher harmonic frequencies when noise was added, resulting in different strain ratings.

Implications for Clinical Evaluation of Strain

The decrease in reliability of strain ratings of samples with noise observed in this study suggests that the auditory-perceptual evaluation of strain may be more challenging for individuals with breathy or dysphonic voices than for individuals with voices without breathiness. The effect of noise on strain may have been the reason that previous studies have observed lower reliability of strain than of other voice qualities (Webb et al., 2004; Zraick et al., 2011) for which individuals with voice disorders were evaluated. When rating strain in dysphonic voices, some listeners may base their ratings more on the presence of noise, whereas other listeners may base their rating more on spectral energies at higher harmonic frequencies or RFF. This finding is similar to the findings of Kreiman et al. (1992), which suggested variability in acoustic cues that individuals use to rate voice quality. Intrarater reliability of strain may have also been low due to noise interacting with other acoustic features of strain. Mid-to-high frequency noise affecting reliability of strain is problematic because many individuals with voice disorders are likely to present increased aspiration noise due to glottal insufficiency. This population needs to be evaluated accurately for effective treatment. Based on our findings, clinicians should be aware that the auditory-perceptual evaluation of strain may not be reliable in individuals with dysphonia and that their strain ratings should be incorporated with care in their clinical practice.

These issues with auditory-perceptual evaluation of strain support call for more research to develop objective measures to assess strain. The findings of this study further support that RFF exhibits potential as an objective measure for assessing increased vocal effort. RFF has consistently differentiated between individuals with healthy voices and vocal hyperfunction (Roy et al., 2016; Stepp et al., 2010, 2011), whereas conventional acoustic measures have been shown mixed results (Belsky et al., in press; Holmberg et al., 2003; Schindler et al., 2013). H1–H2 and measures of spectral tilt (e.g., low-to-high spectral ratio), which reflect increased energies at higher harmonic frequencies, may fail to reflect increased vocal effort if individuals with increased vocal effort do not completely adduct their vocal folds due to structural lesions or vocal fold paralysis. Previous attempts to examine the effect of CPPS on strain have resulted in mixed findings (Anand et al., 2019; Lowell et al., 2012; McKenna & Stepp, 2018), possibly due to occurrences of both increased harmonic energy in higher harmonics (which increases CPPS) and mid-to-high frequency noise (which decreases CPPS) in strained voices. In contrast, RFF is a time-based measure, which is not affected by the spectral contents of voice samples. RFF was also observed to detect possible voice changes from an intense voice-use period that auditory-perception ratings did not reflect (Heller Murray et al., 2016), which suggests that RFF may be more sensitive to small changes in vocal function than auditory-perceptual evaluation. Thus, RFF may be a good complement to clinical evaluation of strain.

Although RFF may reflect strain, the multidimensionality of strain suggests that a single acoustic variable may not be sufficient to capture strain. This study supports assertions about the multidimensional nature of strain, finding statistically significant contributions of RFF and mid-to-high frequency noise to strain in addition to the previously observed effects of increased energies at higher harmonic frequencies (Anand et al., 2019; Bergan et al., 2004). Due to these acoustic features affecting strain, previous studies may have struggled to find a single acoustic measure that correlates strongly with strain (Bhuta et al., 2004) This multidimensional character of strain is also likely to inhibit the recent efforts to develop analogous scales for the perception of voice quality (e.g., *sones* for the loudness scale) from being applied to strain, since developing analogous scales for perception requires a single physical variable that correlates strongly with perception (e.g., noise-to-harmonic ratio for breathiness; Eddins et al., in press).

Instead of a single acoustic measure, multiparametric tools, similar to Acoustic Voice Quality Index (Maryn et al., 2010) and Cepstral Spectral Index of Dysphonia (Awan et al., 2016), may represent strain more adequately. Both Acoustic Voice Quality Index and Cepstral Spectral Index of Dysphonia have been developed to complement clinical evaluation of the overall severity of dysphonia, and a primary acoustic component in both of these indices is CPPS. Although there was a previous attempt to build a multiparametric tool for strain using CPPS as a component (Lowell et al., 2012), CPPS is not likely to specifically represent strain due to both increased energies at higher harmonic and noise frequencies contributing to strain, as previously explained. In order to develop a multiparametric tool for strain, acoustic measures that can independently estimate energies of harmonic and noise frequencies may be required. RFF could also be one of the factors in this tool, with sentences loaded with RFF instances used as speech samples. Future studies should incorporate these acoustic elements and other potential acoustic contributors to strain into a multiparametric tool for strain.

Limitations

Due to our use of synthetically modified samples, listener reactions to any synthetic sound quality in the samples may have affected the results of this study. To determine whether this was the case, we performed an additional VSR of rating synthetic quality, described in Supplemental Material S1. We did not find statistical differences between RFF-modified and unmodified samples, but we did find statistical differences between samples with and without added noise with a large effect size. Increased synthetic quality in the samples with added noise may have affected strain ratings in these samples, but the relationship between synthetic quality and strained voice quality is also unknown. We aimed to add mid-to-high frequency noise as naturally as possible using the breathiness function

in Praat, which estimates spectral shapes of voice samples to filter white noise and generates whispered versions of the original voice samples. However, because we added synthetic noise to natural voice samples, we could not avoid our samples with added noise sounding more synthetic. Future studies should investigate methods to increase noise levels in voice samples more naturally for future perceptual studies of voice quality.

Another limitation of this study is the small number of expert raters in auditory-perceptual evaluations of the original voice recordings and the training stimuli for the VSR task. Because of the known poor interrater reliability of auditory-perceptual evaluation (Webb et al., 2004; Zraick et al., 2011), the scores from these expert raters may have not been reliable. However, these ratings played subsidiary roles of the experiment and thus they are not likely to substantially affect the findings of the study.

Conclusions

Synthetic modification of RFF and addition of mid-to-high frequency noise changed the perceived strain of the modified samples. Lowering RFF resulted in increased strain, and raising RFF resulted in decreased strain, consistent with previous findings of the perceptual studies on RFF. Adding mid-to-high frequency noise resulted in increased strain and decreased intra- and interrater reliability of strain. Our findings support the multidimensionality of strain and suggest that future acoustic assessment of strain can be better achieved through multiparametric tools incorporating multiple acoustic features of strain. The decreased intrarater reliability of strain in the samples with noise indicates that the clinical perceptual evaluation of strain in dysphonic voices can be problematic and further supports the need for objective assessment of strain to complement auditory-perceptual evaluation.

Acknowledgments

This work was supported by Grant DC015570 (awarded to Cara E. Stepp) from the National Institute on Deafness and Other Communication Disorders and a Dudley Allen Sargent Research Fund Grant (awarded to Yeonggwang Park) from Boston University. Thanks to Daniel Buckley, Liz Heller Murray, and Tory McKenna for assistance with training set preparation for the visual sort-and-rate task.

References

American Speech-Language-Hearing Association. (2005). *Guidelines for manual pure-tone threshold audiometry* [Guideline]. https://www.asha.org/policy

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

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., & Roy, N. (2006). Toward the development of an objective index of dysphonia severity: A four-factor acoustic model. *Clinical Linguistics & Phonetics, 20*(1), 35–49. https://doi.org/10.1080/02699200400008353

Awan, S. N., Roy, N., Zhang, D., & Cohen, S. M. (2016). Validation of the cepstral spectral index of dysphonia (CSID) as a screening tool for voice disorders: Development of clinical cutoff scores. *Journal of Voice, 30*(2), 130–144. https://doi.org/10.1016/j.jvoice.2015.04.009

Belsky, M. A., Rothenberger, S. D., Gillespie, A. I., & Gartner-Schmidt, J. L. (in press). Do phonatory aerodynamic and acoustic measures in connected speech differ between vocally healthy adults and patients diagnosed with muscle tension dysphonia? *Journal of Voice*.

Bergan, C. C., Titze, I. R., & Story, B. (2004). The perception of two vocal qualities in a synthesized vocal utterance: Ring and pressed voice. *Journal of Voice, 18*(3), 305–317. https://doi.org/10.1016/j.jvoice.2003.09.004

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. (2019). *Praat: Doing phonetics by computer* (Version 6.0.48). http://www.praat.org

Bowen, L. K., Hands, G. L., Pradhan, S., & Stepp, C. E. (2013). Effects of Parkinson's disease on fundamental frequency variability in running speech. *Journal of Medical Speech-Language Pathology, 21*(3), 235–244.

Brodnitz, F. S. (1966). Rehabilitation of the human voice. *Bulletin of the New York Academy of Medicine, 42*(3), 231–240.

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

Corretge, R. (2019). *Praat Vocal Toolkit*. http://www.praatvocaltoolkit.com

De Bodt, M. S., Van de Heyning, P. H., Wuyts, F. L., & Lambrechts, L. (1996). The perceptual evaluation of voice disorders. *Acta Oto-Rhino-Laryngologica Belgica, 50*(4), 283–291.

Eadie, T. L., Kapsner, M., Rosenzweig, J., Waugh, P., Hillel, A., & Merati, A. (2010). The role of experience on judgments of dysphonia. *Journal of Voice, 24*(5), 564–573. https://doi.org/10.1016/j.jvoice.2008.12.005

Eadie, T. L., & Stepp, C. E. (2013). Acoustic correlate of vocal effort in spasmodic dysphonia. *Annals of Otology, Rhinology & Laryngology, 122*(3), 169–176. https://doi.org/10.1177/000348941312200305

Eddins, D. A., Anand, S., Lang, A., & Shrivastav, R. (in press). Developing clinically relevant scales of breathy and rough voice quality. *Journal of Voice*.

Goberman, A. M., & Blomgren, M. (2008). Fundamental frequency change during offset and onset of voicing in individuals with Parkinson disease. *Journal of Voice, 22*(2), 178–191. https://doi.org/10.1016/j.jvoice.2006.07.006

Granqvist, S. (2003). The visual sort and rate method for perceptual evaluation in listening tests. *Logopedics Phoniatrics Vocology, 28*(3), 109–116. https://doi.org/10.1080/14015430310015255

Heller Murray, E. S., Hands, G. L., Calabrese, C. R., & Stepp, C. E. (2016). Effects of Adventitious acute vocal trauma: Relative fundamental frequency and listener perception. *Journal of Voice, 30*(2), 177–185. https://doi.org/10.1016/j.jvoice.2015.04.005

Heller Murray, E. S., Lien, Y. A., Van Stan, J. H., Mehta, D. D., Hillman, R. E., Noordzij, J. P., & Stepp, C. E. (2017). Relative fundamental frequency distinguishes between phonotraumatic and non-phonotraumatic vocal hyperfunction. *Journal of Speech,*

Language, and Hearing Research, 60(6), 1507–1515. https://doi.org/10.1044/2016_JSLHR-S-16-0262

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

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

Hillman, R. E., Holmberg, E. B., Perkell, J. S., Walsh, M., & Vaughan, C. (1989). Objective assessment of vocal hyperfunction: An experimental framework and initial results. *Journal of Speech and Hearing Research, 32*(2), 373–392. https://doi.org/10.1044/jshr.3202.373

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

Hogikyan, N. D., & Sethuraman, G. (1999). Validation of an instrument to measure voice-related quality of life (V-RQOL). *Journal of Voice, 13*(4), 557–569. https://doi.org/10.1016/S0892-1997(99)80010-1

Holmberg, E. B., Doyle, P., Perkell, J. S., Hammarberg, B., & Hillman, R. E. (2003). Aerodynamic and acoustic voice measurements of patients with vocal nodules: Variation in baseline and changes across voice therapy. *Journal of Voice, 17*(3), 269–282. https://doi.org/10.1067/S0892-1997(03)00076-6

Kawahara, H. (2006). STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds. *Acoustical Science and Technology, 27*(6), 349–353. https://doi.org/10.1250/ast.27.349

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)

Kent, R. D. (1996). Hearing and believing: Some limits to the auditory-perceptual assessment of speech and voice disorders. *American Journal of Speech-Language Pathology, 5*(3), 7–23. https://doi.org/10.1044/1058-0360.0503.07

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *The Journal of Acoustical Society of America, 87*(2), 820–857. https://doi.org/10.1121/1.398894

Kreiman, J., & Gerratt, B. R. (2012). Perceptual interaction of the harmonic source and noise in voice. *The Journal of Acoustical Society of America, 131*(1), 492–500. https://doi.org/10.1121/1.3665997

Kreiman, J., Gerratt, B. R., Kempster, G. B., Erman, A., & Berke, G. S. (1993). Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research. *Journal of Speech and Hearing Research, 36*(1), 21–40. https://doi.org/10.1044/jshr.3601.21

Kreiman, J., Gerratt, B. R., Precoda, K., & Berke, G. S. (1992). Individual differences in voice quality perception. *Journal of Speech and Hearing Research, 35*(3), 512–520. https://doi.org/10.1044/jshr.3503.512

Kreiman, J., Shue, Y. L., Chen, G., Iseli, M., Gerratt, B. R., Neubauer, J., & Alwan, A. (2012). Variability in the relationships among voice quality, harmonic amplitudes, open quotient, and glottal area waveform shape in sustained phonation. *The Journal of Acoustical Society of America, 132*(4), 2625–2632. https://doi.org/10.1121/1.4747007

Lien, Y. S., Michener, C. M., Eadie, T. L., & Stepp, C. E. (2015). Individual monitoring of vocal effort with relative fundamental frequency: Relationships with aerodynamics and listener perception. *Journal of Speech, Language, and Hearing Research, 58*(3), 566–575. https://doi.org/10.1044/2015_JSLHR-S-14-0194

Lowell, S. Y., Kelley, R. T., Awan, S. N., Colton, R. H., & Chan, N. H. (2012). Spectral- and cepstral-based acoustic features of dysphonic, strained voice quality. *Annals of Otology, Rhinology & Laryngology, 121*(8), 539–548. https://doi.org/10.1177/000348941212100808

Macmillan, N. A., & Creelman, D. (2004). *Detection theory: A user's guide* (2nd ed.). Erlbaum.

Maryn, Y., Corthals, P., Van Cauwenberge, P., Roy, N., & De Bodt, M. (2010). Toward improved ecological validity in the acoustic measurement of overall voice quality: Combining continuous speech and sustained vowels. *Journal of Voice, 24*(5), 540–555. https://doi.org/10.1016/j.jvoice.2008.12.014

McKenna, V. S., & Stepp, C. E. (2018). The relationship between acoustical and perceptual measures of vocal effort. *The Journal of Acoustical Society of America, 144*(3), 1643. https://doi.org/10.1121/1.5055234

Morrison, M. D. (1997). Pattern recognition in muscle misuse voice disorders: How I do it. *Journal of Voice, 11*(1), 108–114. https://doi.org/10.1016/S0892-1997(97)80031-8

Oates, J. (2009). Auditory-perceptual evaluation of disordered voice quality: Pros, cons and future directions. *Folia Phoniatrica et Logopaedica, 61*(1), 49–56. https://doi.org/10.1159/000200768

Park, Y., Perkell, J. S., Matthies, M. L., & Stepp, C. E. (2019). Categorization in the perception of breathy voice quality and its relation to voice production in healthy speakers. *Journal of Speech, Language, and Hearing Research, 62*(10), 3655–3666. https://doi.org/10.1044/2019_JSLHR-S-19-0048

Portney, L. G., & Watkins, M. P. (2000). *Foundations of clinical research: Applications to practice*. Prentice Hall.

Ramig, L. O., & Verdolini, K. (1998). Treatment efficacy: voice disorders. *Journal of Speech, Language, and Hearing Research, 41*(1), S101–S116. https://doi.org/10.1044/jslhr.4101.s101

Roy, N. (2003). Functional dysphonia. *Current Opinion in Otolaryngology & Head and Neck Surgery, 11*(3), 144–148. https://doi.org/10.1097/00020840-200306000-00002

Roy, N. (2008). Assessment and treatment of musculoskeletal tension in hyperfunctional voice disorders. *International Journal of Speech-Language Pathology, 10*(4), 195–209. https://doi.org/10.1080/17549500701885577

Roy, N., Barkmeier-Kraemer, J., Eadie, T., Sivasankar, M. P., Mehta, D., Paul, D., & Hillman, R. (2013). Evidence-based clinical voice assessment: A systematic review. *American Journal of Speech-Language Pathology, 22*(2), 212–226. https://doi.org/10.1044/1058-0360(2012/12-0014)

Roy, N., Fetrow, R. A., Merrill, R. M., & Dromey, C. (2016). Exploring the clinical utility of relative fundamental frequency as an objective measure of vocal hyperfunction. *Journal of Speech, Language, and Hearing Research, 59*(5), 1002–1017. https://doi.org/10.1044/2016_JSLHR-S-15-0354

Roy, N., Merrill, R. M., Gray, S. D., & Smith, E. M. (2005). Voice disorders in the general population: Prevalence, risk factors, and occupational impact. *The Laryngoscope, 115*(11), 1988–1995. https://doi.org/10.1097/01.mlg.0000179174.32345.41

Schindler, A., Mozzanica, F., Maruzzi, P., Atac, M., De Cristofaro, V., & Ottaviani, F. (2013). Multidimensional assessment of vocal changes in benign vocal fold lesions after voice therapy. *Auris Nasus Larynx, 40*(3), 291–297. https://doi.org/10.1016/j.anl.2012.08.003

Smith, E., Verdolini, K., Gray, S., Nichols, S., Lemke, J., Barkmeier, J., Dove, H., & Hoffman, H. (1996). Effect of voice disorders on

quality of life. *Journal of Medical Speech-Language Pathology, 4*(4), 223–244.

Stemple, J. C. & Hapner, E. R. (2019). *Voice therapy: Clinical case studies* (5th ed.). Plural.

Stemple, J. C., Roy, N., & Klaben, B. K. (2014). *Clinical voice pathology* (5th ed.). Plural.

Stepp, C. E. (2013). Relative fundamental frequency during vocal onset and offset in older speakers with and without Parkinson's disease. *The Journal of Acoustical Society of America, 133*(3), 1637–1643. https://doi.org/10.1121/1.4776207

Stepp, C. E., Hillman, R. E., & Heaton, J. T. (2010). The impact of vocal hyperfunction on relative fundamental frequency during voicing offset and onset. *Journal of Speech, Language, and Hearing Research, 53*(5), 1220–1226. https://doi.org/10.1044/1092-4388(2010/09-0234)

Stepp, C. E., Merchant, G. R., Heaton, J. T., & Hillman, R. E. (2011). Effects of voice therapy on relative fundamental frequency during voicing offset and onset in patients with vocal hyperfunction. *Journal of Speech, Language, and Hearing Research, 54*(5), 1260–1266. https://doi.org/10.1044/1092-4388(2011/10-0274)

Stepp, C. E., Sawin, D. E., & Eadie, T. L. (2012). The relationship between perception of vocal effort and relative fundamental frequency during voicing offset and onset. *Journal of Speech, Language, and Hearing Research, 55*(6), 1887–1896. https://doi.org/10.1044/1092-4388(2012/11-0294)

Stevens, K. N. (1977). Physics of laryngeal behavior and larynx modes. *Phonetica, 34*(4), 264–279. https://doi.org/10.1159/000259885

Sundberg, J., & Gauffin, J. (1978). Waveform and spectrum of the glottal voice source. *Speech, Music and Hearing Quarterly Progress and Status Report, 19,* 35–50.

Van Stan, J. H., Mehta, D. D., Ortiz, A. J., Burns, J. A., Toles, L. E., Marks, K. L., Vangel, M., Hron, T., Zeitels, S., & Hillman, R. E. (2020). Differences in weeklong ambulatory vocal behavior between female patients with phonotraumatic lesions and matched controls. *Journal of Speech, Language, and Hearing Research, 63*(2), 372–384. https://doi.org/10.1044/2019_JSLHR-19-00065

Verdolini, K., Titze, I. R., & Fennell, A. (1994). Dependence of phonatory effort on hydration level. *Journal of Speech and Hearing Research, 37*(5), 1001–1007. https://doi.org/10.1044/jshr.3705.1001

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. https://doi.org/10.1016/j.jvoice.2016.09.012

Webb, A. L., Carding, P. N., Deary, I. J., MacKenzie, K., Steen, N., & Wilson, J. A. (2004). The reliability of three perceptual evaluation scales for dysphonia. *European Archives of Oto-Rhino-Laryngology, 261*(8), 429–434. https://doi.org/10.1007/s00405-003-0707-7

Zhang, Z. (2015). Regulation of glottal closure and airflow in a three-dimensional phonation model: Implications for vocal intensity control. *The Journal of Acoustical Society of America, 137*(2), 898–910. https://doi.org/10.1121/1.4906272

Zraick, R. I., Kempster, G. B., Connor, N. P., Thibeault, S., Klaben, B. K., Bursac, Z., Thrush, C. R., & Glaze, L. E. (2011). Establishing validity of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V). *American Journal of Speech-Language Pathology, 20*(1), 14–22. https://doi.org/10.1044/1058-0360(2010/09-0105)