JSLHR

Research Article

Test–Retest Reliability of Relative Fundamental Frequency and Conventional Acoustic, Aerodynamic, and Perceptual Measures in Individuals With Healthy Voices

Yeonggwang Park[a] and Cara E. Stepp[a,b,c]

Purpose: Recent studies have shown that an acoustic measure, relative fundamental frequency (RFF), has potential for the assessment of excessive laryngeal tension and vocal effort associated with functional and neurological voice disorders. This study presents an analysis of the test–retest reliability of RFF in individuals with healthy voices and a comparison of reliability between RFF and conventional measures of voice.
Method: Acoustic and aerodynamic measurements and Consensus Auditory–Perceptual Evaluation of Voice (CAPE-V) were performed on 28 individuals with healthy voices on 5 consecutive days. Participants produced RFF stimuli, a sustained /ɑ/, and a reading passage to allow for extraction of acoustic measures and CAPE-V ratings; /pa/ trains were produced to allow for extraction of aerodynamic measures.
Results: Moderate reliabilities (intraclass correlation coefficient [ICC] = .64–.71) were found for RFF values. Mean vocal fundamental frequency, smoothed cepstral peak prominence, shimmer, harmonics-to-noise ratio, and mean airflow rate exhibited good-to-excellent reliabilities (ICC = .76–.99). ICCs for jitter and phonation threshold pressure were moderately reliable (ICC = .67–.74). ICCs for subglottal pressure estimates and all CAPE-V parameters showed poor reliabilities (ICC = .31–.58).
Conclusion: RFF has comparable reliability to conventional measures of voice. This expands the potential for clinical application of RFF.
Supplemental Material: https://doi.org/10.23641/asha.8233376

One of the most common features of voice disorders is vocal hyperfunction (Stemple, Roy, & Klaben, 2014). *Vocal hyperfunction*, characterized by strained voice quality, has been defined as "abuse and/or misuse of the vocal mechanism due to excessive and/or 'imbalanced' muscular forces" (Hillman, Holmberg, Perkell, Walsh, & Vaughan, 1989). Vocal hyperfunction may accompany voice disorders that change the structure of the vocal folds (e.g., vocal nodules), or it may appear in individuals without organic changes to the larynx, as in muscle tension dysphonia (Hillman et al., 1989). Thus, managing vocal hyperfunction is an important therapeutic strategy to treat a variety of voice disorders. However, clinical assessment currently lacks objective tools to quantify the degree of vocal hyperfunction and evaluate the treatment outcomes (Hillman, Gress, Hargrave, Walsh, & Bunting, 1990).

Recently, the acoustic measure relative fundamental frequency (RFF) has been investigated as a possible objective correlate of strained voice quality in vocal hyperfunction (Stepp, Hillman, & Heaton, 2010). RFF quantifies changes in the fundamental frequency (f_0) of voicing offset and onset during the production of sonorant–voiceless consonant–sonorant constructs. In healthy voices, f_0 usually decreases slightly before the voiceless consonant and increases immediately after (Watson, 1998). This f_0 change in voicing offset and onset surrounding voiceless consonants is assumed to be related to an increase in vocal fold tension produced by the cricothyroid muscle, which is thought to aid voicing

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

Editor-in-Chief: Julie Liss
Editor: Jack Jiang
Received December 20, 2018
Revision received February 14, 2019
Accepted February 18, 2019
https://doi.org/10.1044/2019_JSLHR-S-18-0507

Disclosure: *The authors have declared that no competing interests existed at the time of publication.*

Downloaded from: https://pubs.asha.org Boston University on 07/22/2019, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

termination for voiceless consonants (Stevens, 1977). Although the exact cause of the instantaneous f_0 change is still unknown, Stepp et al. (2010) hypothesized that baseline rigidity or tension in the larynx in individuals with vocal hyperfunction would decrease the extent of this f_0 change during voiceless consonant production, and thus, the RFF of individuals with vocal hyperfunction would be lower than in those with healthy voices (Stepp et al., 2010).

Several studies have supported RFF's potential to assess vocal hyperfunction. RFF values were significantly lower in participants with vocal hyperfunction (Heller Murray et al., 2017; Roy, Fetrow, Merrill, & Dromey, 2016; Stepp et al., 2010; Stepp, Sawin, & Eadie, 2012), Parkinson's disease (Bowen, Hands, Pradhan, & Stepp, 2013; Goberman & Blomgren, 2008; Stepp, 2013), and adductor spasmodic dysphonia (Eadie & Stepp, 2013) compared to the RFF of individuals with healthy voices. In addition, the RFF of individuals with vocal hyperfunction significantly increased toward the RFF values of typical speakers after successful voice therapy sessions. This finding suggested promise for the usefulness of RFF as an outcome measure for voice therapy (Roy et al., 2016; Stepp, Merchant, Heaton, & Hillman, 2011). Studies have also evaluated RFF's ability to assess the degree of baseline laryngeal tension, the findings of which included significant correlations between RFF and both aerodynamic (Lien, Michener, Eadie, & Stepp, 2015) and auditory–perceptual measures (Stepp et al., 2012) of vocal effort, and with a kinematic estimate of laryngeal stiffness (McKenna, Heller Murray, Lien, & Stepp, 2016).

Although RFF continues to show promise as a possible objective marker for vocal hyperfunction, more research is necessary before it can be utilized clinically, such as reliability, sensitivity to change, and diagnostic sensitivity and specificity. This study aimed to examine test–retest reliability of RFF. A few studies have compared RFF values estimated at different times, and no significant group differences in RFF values of healthy individuals were found when measured at 10 weeks apart (Heller Murray, Hands, Calabrese, & Stepp, 2016) and 1 hr apart (Roy et al., 2016). However, very little is known about RFF's reproducibility in individual speakers over time. Group effects, though important, can mask whether a measure is a useful indicator at the individual patient level. We examined the test–retest reliability in individuals with healthy voices to minimize any voice changes that could affect the results. We measured the participants' voices every day throughout one work week during which their vocal function was assumed to be relatively stable.

We also measured participants' voices using conventional voice measures throughout the week in order to compare the test–retest reliability of these standard clinical measures with the reliability of RFF. Instrumental measures have provided valuable information to clinicians when they diagnose voice disorders and assess their severity, prognosis, and treatment outcomes (Stemple et al., 2014). Acoustic measures of mean vocal f_0, jitter, shimmer, and harmonics-to-noise ratio (HNR) were selected because of their frequent usage in clinic, as well as the research evidence that suggests their effectiveness in classifying dysphonia (Desjardins, Halstead, Cooke, & Bonilha, 2017; Eadie & Doyle, 2005; Linder, Albers, Hess, Poppl, & Schonweiler, 2008). Jitter, shimmer, and HNR have been studied due to their hypothesized association with roughness and breathiness (Eskenazi, Childers, & Hicks, 1990; Hillenbrand, 1988). Although the test–retest reliabilities of jitter, shimmer, and HNR have shown mixed findings in previous studies (Bough, Heuer, Sataloff, Hills, & Cater, 1996; Carding et al., 2004; Leong et al., 2013), we included them since they have been commonly used in both clinical and research applications. We also included smoothed cepstral peak prominence (CPPS), a cepstral measure obtained from the Fourier transform of the power spectrum, which has also shown high accuracy in predicting dysphonia (Heman-Ackah et al., 2003). Aerodynamic measures of mean airflow rate, subglottic pressure, and phonation threshold pressure (PTP) that have shown effectiveness in detecting vocal changes (Chang & Karnell, 2004; Desjardins et al., 2017; Solomon & DiMattia, 2000) were included, as well. Comparing the reliability of these conventional measures with that of RFF could help determine the clinical usefulness of RFF relative to those measures that are already in use in clinical practice. In addition, Consensus Auditory–Perceptual Evaluation of Voice (CAPE-V; American Speech-Language-Hearing Association, 2002) was included since perceptual evaluation is routinely used in clinics (Carding, Wilson, MacKenzie, & Deary, 2009; De Bodt, Van de Heyning, Wuyts, & Lambrechts, 1996). We hypothesized that RFF would have comparable but slightly lower reliability than the conventional acoustic and aerodynamic measures because most of them are associated with perceived overall dysphonia, often caused by structural changes of the vocal folds, whereas RFF is thought to reflect strain related to laryngeal tension, a functional aspect of vocal production. We hypothesized that RFF would have higher reliability than CAPE-V ratings because of the subjective nature of perceptual evaluation.

We also examined the effect of speaker intensity on RFF reliability. Controlling speaker intensity levels in different recording sessions produces more reliable results in acoustic and aerodynamic measurements (Lee, Stemple, & Kizer, 1999), since different intensity levels between recording sessions can lead to higher variability. Thus, the reliability of RFF stimuli produced in soft, comfortable, and loud voices was compared with the hypothesis that different loudness levels would result in different reliability. Although Park and Stepp (2018) examined the within-subject standard deviation of different loudness levels and found no effect of loudness, the within-subject standard deviations were estimated from nine RFF values obtained from one recording session, not on different days. In this study, we examined the reliabilities of RFF mean values produced at different loudness levels over five consecutive days. Since a wider range of loudness may be produced relative to the ranges elicited by instructions for comfortable and soft voices, we hypothesized that using a loud voice would result in lower reliability of RFF due to less consistent

Downloaded from: https://pubs.asha.org Boston University on 07/22/2019, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

sound pressure levels over the experimental week than when using soft and comfortable voices.

Method

Participants

Thirty-two healthy participants aged 18–33 years (16 women, 16 men; $M = 22.5$, $SD = 4.1$) were recruited and reported no prior history of speech, language, and hearing disorders. Participants also reported a small-to-medium amount of daily voice use, classified as low voice users. The low voice users were recruited in order to minimize possible voice changes during the week-long course of the study. Participants completed the voice handicap index (VHI) and the reflux symptom index (RSI). VHI and RSI, both self-rating questionnaires, subjectively evaluate the degree of voice handicap and laryngopharyngeal reflux, respectively (Belafsky, Postma, & Koufman, 2002; Jacobson et al., 1997). Participants completed the questionnaires on their first study visit and scored within normal ranges, except for one participant who scored higher than the cutoff score for VHI and was excluded. In addition, all participants 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). The participants provided written consent prior to participation, in compliance with the Boston University Institutional Review Board. Three additional participants were excluded midway through the experimental week due to sickness. Thus, a total of four participants were excluded during the course of the study, resulting in 28 participants.

Experimental Tasks

Participants visited the lab over a period of five consecutive days, Monday through Friday of the same week. They were asked to come at a similar time each day, at least 3 hr after waking, such that their voice conditions would be as consistent as possible during each visit, although there were some cases in which participants had to schedule different times. In order to confirm that their voice conditions were consistent throughout the week, we conducted a detailed voice use interview and a self-administered vocal rating during each visit.

The chronological order of experimental tasks during each visit is outlined in Table 1. RFF stimuli recording was performed after the conventional acoustic measurements to ensure that the loud phonation in the RFF protocol did not have an impact on participants' voices and thus affect the results of later recordings. Aerodynamic measurement, which also had a loud voice condition, was placed after the conventional acoustic and RFF recordings for the same reason.

Table 1. The approximate timeline of experimental tasks during each visit.

Experimental task	Time (min)
Voice interview	2
Vocal self-rating (IPSV)	3
Training of speech stimuli and recording set up	5
Conventional acoustic measurement	3
RFF stimuli	2
Aerodynamic task training and measurement	10

Note. IPSV = Inability to Produce Soft Voice; RFF = relative fundamental frequency.

Voice Interview and Vocal Self-Rating

During every visit, the experimenter interviewed participants with detailed questions about their daily voice use, voice condition, and wake-up time to document any behaviors that might affect their current voice condition. Participants also performed a vocal self-rating task called the *Inability to Produce Soft Voice* (IPSV), which has shown reliability in tracking teachers' voice changes (Halpern, Spielman, Hunter, & Titze, 2009). IPSV consists of four different tasks, which participants are asked to perform as softly as possible: (a) sustaining /i/ for 5 s on a comfortable pitch, (b) gliding on /i/ from a low to a high pitch, (c) saying a train of /i/ production in staccato with a high pitch, and (d) singing a few bars of "Happy Birthday" in a high pitch. After these tasks, participants rated their own score on a scale of 1 (*no problem*) to 10 (*extreme problem*).

Conventional Acoustic Measurements

Participants were equipped with a head-mounted microphone (Shure WH20) in a sound-treated booth. Their voice was recorded with SONAR Artist (Cakewalk) using a 44.1-kHz sampling rate. Participants produced sustained /ɑ/ vowels for 3–5 s in one exhalation with a constant pitch and loudness. We asked participants to produce the sustained utterance nine times to match the sample numbers with RFF productions to more accurately compare the variabilities of conventional measures and RFF. Participants were also asked to read the first paragraph of the "Rainbow Passage" (Fairbanks, 1960).

RFF Stimuli Recording

RFF stimuli were recorded with the same recording equipment. The RFF short utterance stimulus, /əfə/, was selected because it resulted in lower RFF within-subject standard deviation compared to other stimuli in previous studies (Lien, Gattuccio, & Stepp, 2014; Park & Stepp, 2018). In addition, participants were asked to produce RFF stimuli with equal stress using similar pitch and loudness in both vowels since this has also been shown to decrease within-subject standard deviations (Park & Stepp, 2018).

RFF stimuli were produced with comfortable, soft, and loud voices to test the effect of loudness on RFF reliability. The degree of softness and loudness were not assigned a specific sound pressure level but, rather, were determined by participants, similar to clinical instructions for loud voice (Patel et al., 2018). However, the sound pressure level was estimated offline by calibrating the acoustic waveforms collected using an electrolarynx and

Downloaded from: https://pubs.asha.org Boston University on 07/22/2019, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

sound pressure level meter. The participants were asked not to whisper for the soft voice, since accurate estimation of f_0 is difficult with whispered voice. Each stimulus under a given loudness condition was produced nine times, as RFF has been shown to correlate better with auditory–perceptual judgments when averaged over at least six productions (Eadie & Stepp, 2013).

The experimenter instructed the participants on how to produce stimuli for both conventional acoustic measures and RFF before each recording session (the instruction included practice of each task except for the reading of the "Rainbow Passage"). Participants also listened to sample recordings of RFF stimuli to learn how to produce RFF stimuli with the equal stress; opposite gender recordings were played to avoid pitch mimicking. During the recording, when participants occasionally pronounced the stimuli with the wrong stress, the experimenter asked them to produce the stimulus again. The experimenter also asked the participants to repeat any stimulus produced with clear glottalization or with extremely short vowel productions, as these do not allow for calculation of RFF.

Aerodynamic Measurement

Mean airflow rate (ml/s) and intraoral estimates of subglottic pressure (Psub; cm H_2O) were measured with a phonatory aerodynamic system (Model 6600, PENTAX Medical). Participants wore a face mask, which fit over their nose and mouth, and placed a small catheter inside their mouth. They produced six trains of five /pɑ/s with a comfortable loudness level and six trains with a loud level. For PTP, participants produced continuous /pɑ/s in one exhaling breath with decreasing loudness until their voice stopped. They performed this protocol three times at a comfortable pitch. They performed the same protocol three times each at 3 semitones (ST) below and above their comfortable f_0 (Enflo, Sundberg, Romedahl, & McAllister, 2013). Each individual's comfortable f_0 was determined from each participant's recording of sustained /ɑ/ using Praat acoustic software (Boersma & Weenink, 2016). Sample synthetic voices at each participant's comfortable f_0 and 3 ST below and above their comfortable f_0 were generated and played with a Madde synthesizer (Granqvist, 2010). We played each sample to the participants, and they produced the /pa/s with the f_0 they heard. The participants were given instructions and practiced before each task.

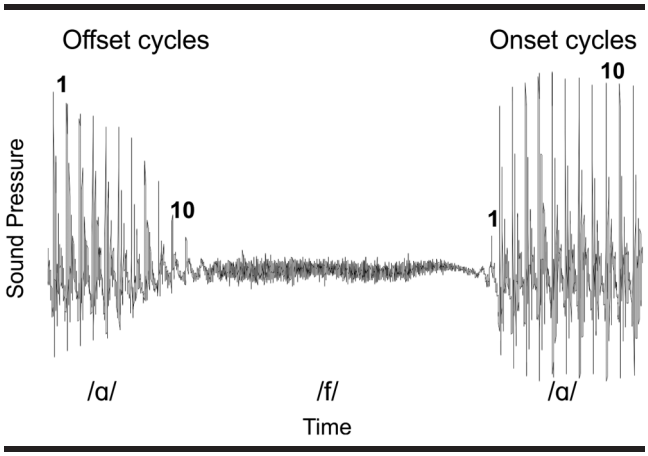
Data Analysis

RFF was estimated using an automated RFF estimation algorithm (Lien et al., 2017) in MATLAB (Version R2015b, MathWorks). The automated RFF algorithm identified the voiced cycles before and after the consonants, estimated the periods and the instantaneous f_0 for each cycle, and calculated RFF with the RFF equation $ST = 39.86 \times \log_{10}(f_0 / \text{reference } f_0)$. The algorithm automatically rejected any recorded stimuli that lack periodic cycles or contain glottalization, which can affect the accuracy of the f_0 estimation. We focused on Offset Cycle 10 and Onset Cycle 1 RFF (see Figure 1) values because these two cycles best represent the degree of vocal hyperfunction (Lien et al., 2015; Stepp et al., 2010). The mean Offset Cycle 10 and Onset Cycle 1 RFF values were calculated with the nine RFF values estimated (less in the case of any rejections) from the nine recordings of each stimulus condition.

Conventional acoustic measures were obtained using Praat acoustic analysis software (Version 6.0.21). Praat's built-in function, "voice report," was used to estimate acoustic measures for a selected stable 1-s segment of the recorded sustained vowel /ɑ/ samples. Jitter, shimmer, and HNR were obtained from the nine sustained /ɑ/ productions, and the nine values of each measure on a given experimental day were averaged to obtain the mean value of each individual parameter for the day. CPPS was also obtained from a Praat built-in function developed by Maryn and Weenink (2015), following the protocol in Watts, Awan, and Maryn (2016). Nine CPPS values were obtained from each of the nine 1-s sustained /ɑ/ recordings and averaged into $CPPS_{vowel}$ for each day. Mean vocal f_0 and $CPPS_{sentence}$ were calculated from the recordings of the first and second sentences of the "Rainbow Passage" (Fairbanks, 1960).

Aerodynamic data were analyzed in MATLAB to obtain mean airflow rate and subglottic pressure (see Supplemental Material S1 for the analysis scripts). Airflow rate and intraoral air pressure signals were extracted from the raw aerodynamic data from the phonatory aerodynamic system. In order to be considered a valid measurement, the airflow signal had to contain a steady-state, horizontal portion during the vowel, and the air pressure signal had to have a flat peak during /p/ stop consonant (Patel et al., 2018). From the six /pɑ/ trains of each airflow rate and air pressure signals, three stable /pɑ/ trains were chosen. In each selected /pɑ/ train, measures were obtained from the middle three syllables, discarding the first and last syllables. In airflow rate signals, the horizontal portions during the vowels were selected, and nine selections from three /pa/ trains (3 middle vowels × 3 /pa/ trains) were averaged into

Figure 1. An example acoustic waveform of a sonorant–voiceless consonant–sonorant. The Offset Cycles 1 and 10 and Onset Cycles 1 and 10 are labeled.



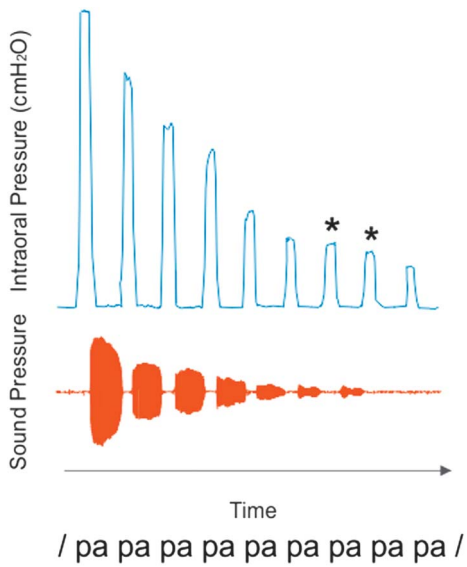
Downloaded from: https://pubs.asha.org Boston University on 07/22/2019, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

mean airflow rate for each day. In intraoral air pressure signals, the interpolation method (Patel et al., 2018) was used to estimate the subglottic pressure over the vowels, and nine estimate subglottic pressures (3 middle vowels × 3 /pa/ trains) were averaged into Psub for each day. Because oral estimates of subglottal pressure are known to be underestimates if the pressure in the oral cavity is not equalized (Fryd, Van Stan, Hillman, & Mehta, 2016), we decided to eliminate more "peaky" pressure waveforms by setting a threshold value to the 5% variation of each pressure peak, described in McKenna et al. (2016), and eliminating pressure waveforms with variation above this threshold value from the estimation. We eliminated approximately 10% of all pressure waveforms. However, we did not see differences in the results based on whether these waveforms were included, consistent with the finding of Fryd et al. (2016).

PTP was estimated in MATLAB (see Figure 2). PTP was estimated as the average of the two intraoral pressure peaks that were surrounding where the acoustic waveform indicated that the voice of the participant stopped or turned into a whisper (Enflo et al., 2013). This selection was performed on all three trials at each f_0, and the selected values were averaged across f_0 to represent the PTP for that day.

CAPE-V was performed by a voice-experienced speech pathologist. The total number of the ratings were 140 (28 participants × 5 days), and the order of the ratings was pseudorandomized across both participants and days. The listening samples consisted of the three 1-s /ɑ/s and the first two sentences of the recording of the "Rainbow Passage" from the acoustic recordings of each participant on each day. After listening to each sample, the rater completed the standardized CAPE-V form that contained 100-mm visual analog scales (Kempster, Gerratt, Verdolini Abbott, Barkmeier-Kraemer, & Hillman, 2009). Because all of our participants had healthy voices, we excluded pitch and loudness parameters in the CAPE-V and examined the four parameters of voice quality: overall severity, roughness, breathiness, and strain. Because of the large number of ratings, the rater completed the ratings over five different sessions. For intrarater reliability evaluation, 20% of the ratings were repeated by the same rater at a later date, and for interrater reliability evaluation, another speech-language pathologist with experience in voice also performed 20% of the ratings.

Figure 2. Example of phonation threshold pressure estimation. The top panel is the intraoral pressure. The bottom panel is the associated acoustic waveform. Two peaks (marked with asterisks) were selected because they surrounded the point at which the voice stopped (phonation threshold).



Statistical Analysis

To compare measurement variabilities within healthy speakers, intraclass correlation coefficients (ICCs) were used in previous studies (Awan, Novaleski, & Yingling, 2013; Bough et al., 1996; Carding et al., 2004; Leong et al., 2013). Thus, ICC values and their 95% confidence intervals for each acoustic and aerodynamic measure and each CAPE-V parameter were calculated using SPSS (Version 24, SPSS, Inc.) based on single measures, absolute agreement, and the two-way mixed-effects model (McGraw & Wong, 1996). Although there are no standards to interpret ICCs, ICCs below .5 have been suggested as indicative of poor reliability, ICCs of .5–.75 as moderately reliable, ICCs of .75–.9 as having good reliability, and ICCs above .9 as having excellent reliability (Koo & Li, 2016). In order to assess intrarater reliabilities for the CAPE-V parameters, Pearson's correlation coefficients (r) were calculated. Interrater reliabilities for the CAPE-V parameters were calculated with ICCs as the two-way mixed-effects model for the consistency of single measurements for each CAPE-V parameter. Intrarater and interrater reliabilities for each parameter are presented in Table 2. Intrarater reliabilities were ≥ .64 for all CAPE-V parameters, except strain (r = .34), and interrater reliabilities were poor ($r \leq .27$) for all parameters.

As a post hoc assessment, we also examined the possible effects of the participants' time awake (time between their wake-up time and the recording session) on their voices. Most of the recording sessions were scheduled at a similar time of the day during the week for each participant, but some participants had to schedule at a different time of

Table 2. Intrarater and interrater reliability of Consensus Auditory–Perceptual Evaluation of Voice parameters.

Parameters	Intrarater Pearson *r*	Interrater ICC
Overall severity	.79	.21
Roughness	.57	.21
Breathiness	.64	.00
Strain	.34	.27

Note. ICC = intraclass correlation coefficient.

Downloaded from: https://pubs.asha.org Boston University on 07/22/2019, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

the day or woke up much earlier or later than the other days. We suspected that being awake much longer and possibly talking more before the recording session may have affected these participants' voices. We chose 10 participants whose ranges of the time awake varied by more than 5 hr across the experimental week, so that the sample would have sufficient variance in the associated measures to show potential associations. For each of the 10 participants, individual Pearson correlation analysis was performed between the time awake, mean vocal f_0, Offset 10 and Onset 1 RFF values, and PTP from the 5-day sessions. These instrumental measures were specifically chosen because of their known sensitivities to vocal loading or vocal fatigue (Chang & Karnell, 2004; Kagan & Heaton, 2017; Solomon & DiMattia, 2000; Stemple, Stanley, & Lee, 1995; Welham & Maclagan, 2003). The individual Pearson correlation coefficients (r) from the 10 participants were averaged using Fisher's z' transformation.

Results

Mean values of all of the CAPE-V parameters are presented in Table 3. All of the parameters in the CAPE-V showed low mean values and thus support that participants had healthy voices. All of our participants had Type 1 voices as determined by the first author.

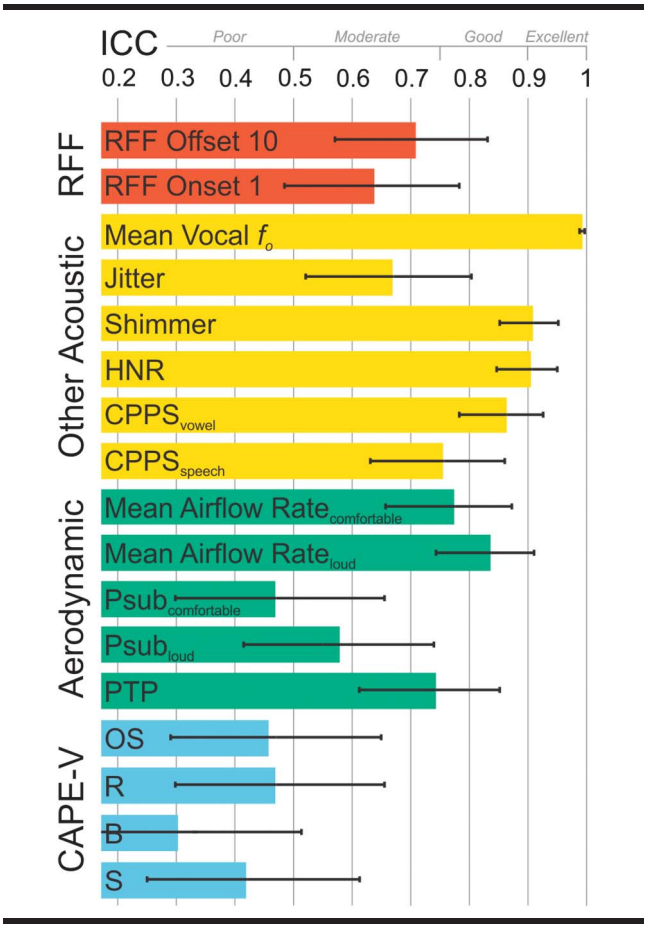
We obtained ICCs for Offset 10 RFF and Onset 1 RFF as well as for conventional acoustic, aerodynamic, and perceptual measures from the recordings of five consecutive days in order to assess the test–retest reliability of RFF. The results are presented in Figure 3. Both Offset 10 and Onset 1 RFF had moderate reliability. Excellent reliability was observed for mean vocal f_0, shimmer, and HNR, and good reliability was seen for both $CPPS_{vowel}$ and $CPPS_{speech}$. For the aerodynamic measures, both mean airflow rates measured in both comfortable and loud voice showed good reliability, PTP showed moderate-to-good reliability, and both $Psub_{comfortable}$ and $Psub_{loud}$ showed poor-to-moderate reliability. All of the CAPE-V parameters exhibited poor reliability.

We also obtained the ICCs for RFF from different loudness levels (see Figure 4) in order to compare the effects of loudness on the test–retest reliability. Analysis of soft voice recordings produced the highest ICC for Onset 1 RFF, raising the ICC to "good" reliability. Loud voice recordings had good reliability in Offset 10 values, but poor reliability in Onset 1 values.

Table 3. Mean values of Consensus Auditory–Perceptual Evaluation of Voice parameters.

Parameters	*M* (*SD*)
Overall severity	4.6 (2.7)
Roughness	2.0 (2.3)
Breathiness	1.8 (1.9)
Strain	1.2 (1.6)

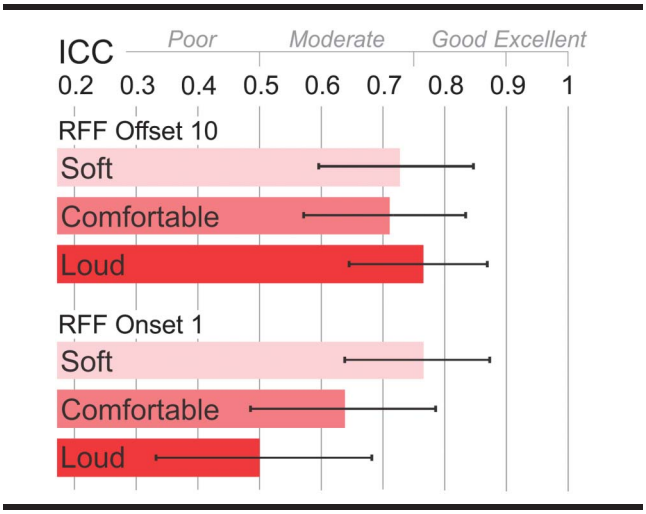
Figure 3. Intraclass correlation coefficient (ICC) values obtained with five measurements over five consecutive days (error bars indicate the 95% confidence intervals). ICCs below .5 are considered poor, ICCs of .5–.75 are moderately reliable, ICCs of .75–.9 are good reliability, and ICCs above .9 are excellent reliability. ICCs of relative fundamental frequency (RFF) are shown as red, other acoustic measures are shown as yellow, aerodynamic measures are shown as green, and Consensus Auditory–Perceptual Evaluation of Voice (CAPE-V) parameters are shown as blue. f_0 = fundamental frequency; HNR = harmonic-to-noise ratio; CPPS = smoothed cepstral peak prominence; Psub = subglottic pressure estimate; PTP = phonation threshold pressure; OS = overall severity; R = roughness; B = breathiness; S = strain.



The results of the voice-related daily questionnaires, as well as the daily interviews and IPSV scores, suggest that participants' voices did not change during the experimental week. Participants had normal VHI scores (M = 8.2, SD = 7.3) and RSI scores (M = 3.3, SD = 3.2). Participants did not report any significant voice use with the exception of two participants who reported yelling during a sporting event. Most participants also did not report any discomfort in the throat, with the exception of a few participants reporting slight laryngeal discomfort during some of the recording sessions. The mean IPSV value collected from participants was 2.3, and the mean within-subject standard deviation of IPSV was 0.6.

Downloaded from: https://pubs.asha.org Boston University on 07/22/2019, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

Figure 4. Intraclass correlation coefficients (ICCs; error bars indicate 95% confidence intervals) of relative fundamental frequency (RFF) produced with different loudness levels.



The log of participants' wake-up times and recording session times indicated that some participants had substantial differences in their wake-up or recording session times of over 5 days. The average range of the time awake before the recording sessions (duration between wake-up time and recording session) over the 5 days was 3.5 hr, and 10 participants had ranges of their time awake that were greater than 5 hr. The possible effects of this variability in time awake before the recording session on participants' voices were evaluated using averaged individual Pearson correlation coefficients (see Table 4). The only measures with an averaged correlation coefficient greater than ±.5 (moderate correlation) with the time awake were Onset 1 RFF (−.50) and mean vocal f_0 (.55; see Table 4).

Discussion

ICC for RFF Versus Other Measures

The test–retest reliabilities of RFF and conventional acoustic, aerodynamic, and perceptual measures were assessed using ICC values (see Figure 3). Both Offset 10 and Onset 1 RFF were found to be moderately reliable, which were lower than the reliabilities of mean vocal f_0, shimmer, HNR, $CPPS_{vowel}$, mean airflow $rate_{comfortable}$, and mean airflow $rate_{loud}$. The lower ICCs for RFF compared to ICCs for these conventional measures were expected because RFF measures are thought to correlate with laryngeal tension, whereas most of the measures that showed higher reliabilities than RFF, except mean vocal f_0, correlate with overall dysphonia severity; thus, RFF is likely to be more sensitive to day-to-day functional variation than these measures. The ICCs for RFF were similar to the ICC for PTP, possibly because PTP also reflects day-to-day variations in vocal fold vibratory characteristics and vocal fatigue (Chan & Titze, 2006; Solomon & DiMattia, 2000).

Another potential reason that the test–retest reliability of RFF was lower than that of many conventional instrumental measures could be the difference in the speech samples. RFF is measured in a continuous speech context, but most of the conventional measures are measured during sustained phonation. Sustained phonation may result in better test–retest reliability than continuous speech because it is free from fluctuations in frequency and amplitude due to prosody and is not affected by speech rate (Maryn, Corthals, Van Cauwenberge, Roy, & De Bodt, 2010). Leong et al. (2013) observed higher ICC values in voice quality measures obtained from sustained vowels compared to the measures obtained from sentence stimuli. Similarly, in our study, $CPPS_{speech}$, measured from sentence stimuli, had lower test–retest reliability than $CPPS_{vowel}$, which was measured from a sustained /ɑ/ phonation. The ICC for $CPPS_{speech}$, in fact, was similar to the ICC value for RFF. The lower test–retest reliability of RFF compared to other conventional instrumental measures may be, in part, the result of differences in stimuli.

The test–retest reliability of RFF was higher than the test–retest reliabilities of CAPE-V parameters, as expected. The test–retest reliabilities of CAPE-V parameters were poor (below .5), and these results might be due to low intrarater and interrater reliabilities. Intrarater reliabilities for roughness and breathiness (see Table 2) were below the averaged intrarater reliabilities (roughness: r = .77, breathiness: r = .82) obtained from 21 voice-trained speech-language pathologists (Zraick et al., 2011). In addition, both intrarater reliabilities of strain in our study (see Table 2) and in the previous study (r = .35; Zraick et al., 2011) were poor to moderate, despite that the speech-language pathologist was an expert in voice with experience in administering the

Table 4. Averaged individual correlation coefficients (r) and standard errors between the time awake, Offset 10 and Onset 1 relative fundamental frequency (RFF), mean vocal fundamental frequency (f_0), and phonation threshold pressure (PTP) among 10 participants whose range of the time awake during the experimental week was over 5 hr.

Measures	Time awake	Offset 10 RFF	Onset 1 RFF	Mean vocal f_0	PTP
Time awake	1.00	−.44 (.21)	−.50 (.16)	.55 (.22)	−.09 (.21)
Offset 10 RFF		1.00	.39 (.18)	−.44 (.21)	−.21 (.21)
Onset 1 RFF			1.00	−.41 (.17)	.09 (.19)
Mean vocal f_0				1.00	−.11 (.22)
PTP					1.00

Downloaded from: https://pubs.asha.org Boston University on 07/22/2019, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

CAPE-V, which has shown to increase intrarater reliability (Eadie & Baylor, 2006). Auditory fatigue may be one of the reasons for poor intrarater reliabilities, although we aimed to minimize the potential for fatigue by dividing the rating sessions into five different days. Having only one rater might have also resulted in poor test–retest reliabilities (5-day data), since a previous study with seven raters resulted in good test–retest reliabilities for a similar perceptual rating, the Grade, Roughness, Breathiness, Asthenia, and Strain Scale (Webb et al., 2004). However, in typical clinical settings, only one speech-language pathologist is likely to perform the CAPE-V, and thus, our results might be a more accurate reflection of the actual test–retest reliability of CAPE-V in practice. Nevertheless, the test–retest reliabilities of CAPE-V performed by one rater is likely to be heavily dependent on who the rater is, as suggested by the poor interrater reliabilities (see Table 2). The interrater reliabilities of all CAPE-V parameters in our study were lower than published values with 21 raters (Zraick et al., 2011). These results confirm the need for more objective measures, in addition to perceptual measures, in clinical settings (Hillman et al., 1990). The higher reliabilities of RFF compared to the reliabilities of CAPE-V, especially the strain parameter, support the potential clinical utility of RFF, since perceptual evaluation is widely used for assessing hyperfunctional voice disorders and evaluating treatment outcomes (Carding et al., 2009).

Soft Voice in RFF Test–Retest Reliability

We included different loudness levels in RFF stimuli recordings to see if loudness would affect the test–retest reliability of RFF. We found that loudness did not affect the reliability of Offset 10 values, but the use of loud voice decreased the reliability for Onset 1 RFF, whereas the soft voice increased the reliability for Onset 1 values (see Figure 4). This is somewhat surprising because soft voice has been associated with both increased values and variabilities of jitter and shimmer, which may reflect increased variability in vocal fold vibratory characteristics (Brockmann, Storck, Carding, & Drinnan, 2008). Soft voice has also been associated with vocal fatigue, as it is used in tasks for IPSV and PTP (Chan & Titze, 2006; Hunter, 2011; Solomon & DiMattia, 2000). We also previously found that RFF values produced with a soft voice had high between-subjects variability (Park & Stepp, 2018). One possible explanation for the higher ICC value for soft voice in the current study may be that participants used more consistent vocal effort to produce soft voice than comfortable and loud voice, and RFF may be sensitive to this vocal effort.

Comparison to the Literature

The observed ICCs for most of the measures in our study were generally higher compared to those reported in previous studies (see Table 5), possibly suggesting that the participants' voices were less varied during the experiment in our study. Four previous studies have documented ICC values for the test–retest reliabilities of the measures included in the current study. Bough et al. (1996) examined the test–retest reliabilities of mean vocal f_0, jitter, shimmer, and HNR over 15 test sessions at consistent times of the day. Leong et al. (2013) evaluated the test–retest reliabilities of mean vocal f_0, jitter, shimmer, and CPPS over 10 sessions. Carding et al. (2004) evaluated the test–retest reliabilities of jitter, shimmer, and HNR in 45 participants over 2 hr. Awan et al. (2013) examined the test–retest reliabilities of aerodynamic measures over two sessions. The results from the previous studies were within the 95% confidence intervals of the results from the current study for jitter, HNR, $CPPS_{speech}$, and mean airflow rate (bolded; see Table 5). For acoustic measures in general, we observed higher reliability in our study, and we suspect that having sessions over consecutive days may have resulted in more consistent vocal conditions between the recording sessions compared to other studies. Although there are also other factors that could have affected the reliabilities, including recording environment (Deliyski, Shaw, Evans, & Vesselinov, 2006), gender distributions, and the number of participants and sessions, the higher reliabilities of the acoustic measures in our study may suggest that the participants' voices were more consistent during the experimental week, which may be a better environment to assess reliabilities of instrumental measures.

On the other hand, the ICC for Psub was generally lower in our study. We suspect that low ICCs of Psub may have been from varied intensity producing Psub tasks and thus normalized Psub with decibels of sound pressure level, as described in the study of Espinoza, Zanartu, Van Stan, Mehta, and Hillman (2017), and recalculated ICCs; however, we obtained similar ICCs using normalization (comfortable: .51, loud: .61), suggesting that the low ICCs of Psub are not due to intensity variability. Another possible reason for the low reliability may be that the Psub measurement contained peaky pressure waveforms, which are known to result in underestimation (Holmberg, Perkell, & Hillman, 1984). However, as mentioned in the Method section, we eliminated some particularly peaky pressure waveforms and found no differences in ICC values; there is still a possibility that our data include peaky pressure waveforms that could have resulted in underestimation. Finally, we found that many of our participants produced /pa/ trains at a slower rate than the recommended rate of 1.5–2 /pa/s per second, which might have led to more peaky pressure waveforms (Holmberg et al., 1984).

Time Awake Versus the Outcome Measures

We examined the times between the participants' wake-up time and the recording sessions because we suspected that this time difference may affect the voice condition during the experimental week. We found that this time awake had moderate correlations with Onset 1 RFF ($r = -.50$) and mean vocal f_0 ($r = .55$). We hypothesize that, when the participants were awake longer before the session, they were likely to have talked more prior to the session, and the

Downloaded from: https://pubs.asha.org Boston University on 07/22/2019, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

Table 5. Comparison of test–retest reliability intraclass correlation coefficient (ICC) values to the literature.

Category	Measures	Current study's ICC	Previous studies' ICCs
Acoustic measures	Mean vocal f_0	.99	.32 (female), .60 (male)[a]
	Jitter	**.67**	.50 (female), .91 (male)[a] .32[b], **.73**[b]
	Shimmer	.91	.56 (female), .53 (male)[a] .67[b], .55[c]
	HNR	.91	.23 (female), .05 (male)[a] **.93**[b], .68[c]
	$CPPS_{speech}$	.76	.45 (female), **.80 (male)**[a]
Aerodynamic measures	Mean airflow rate	.78 (comfortable) .84 (loud)	**.67 (comfortable)**[d]
	Subglottic pressure	.47 (comfortable) .58 (loud)	.74 (comfortable)[d]

Note. Bolded are the ICC values from the previous studies that were within 95% confidence intervals of the ICC values from the current study. f_0 = fundamental frequency; HNR = harmonics-to-noise ratio; CPPS = cepstral peak prominence.
[a]Leong et al. (2013). [b]Bough et al. (1996). [c]Carding et al. (2004). [d]Awan et al. (2013).

increased vocalization prior to the session might have increased their baseline laryngeal tension; both RFF and mean vocal f_0 may reflect this change. This finding is similar to the finding of Garrett and Healey (1987), who measured participants' voices three times during a day and found that male participants showed a significant increase in their mean vocal f_0 at the later times of the day. This increase in mean vocal f_0 was consistent with the findings of Stemple et al. (1995), who observed significant increases in mean vocal f_0 in both sustained vowel and reading samples after vocal loading tasks (Stemple et al., 1995). However, previous studies have mixed findings about the effect of vocal loading tasks on RFF (Fujiki, Chapleau, Sundarrajan, McKenna, & Sivasankar, 2017; Kagan & Heaton, 2017), and PTP, a sensitive measure to vocal loading tasks (Chang & Karnell, 2004; Solomon & DiMattia, 2000), was not correlated with the time awake ($r = –.09$) in the current study. Thus, the time awake may not modulate only the amount of vocalization before the sessions, but it may have influenced other possible factors that could have affected RFF and mean vocal f_0, but not PTP. Because the time awake was shown to be correlated with RFF, the difference in the range of the time awake during the experimental week may have resulted in the moderate test–retest reliabilities of RFF. In contrast, the test–retest reliability of mean vocal f_0 was excellent. Mean vocal f_0 may be less sensitive to vocal change due to the time awake, thus RFF, which is more likely to be related to the baseline laryngeal tension (Lien et al., 2015; McKenna et al., 2016).

Limitations and Future Directions

We examined the test–retest reliability of clinical voice outcome measures, recruiting individuals who reported no history of voice disorders. However, they were not examined by a laryngologist. If they had some degree of vocal hyperfunction or other voice disorder, ICC results and the acoustic measures would not represent solely individuals with healthy voices. The test–retest reliabilities of acoustic measures have shown to be less reliable in dysphonic voices compared to healthy voices (Carding et al., 2004). Individuals with dysphonic voices would have more irregular voice conditions than individuals with healthy voices; thus, test–retest reliabilities of the measures among dysphonic voices should be examined in the future. In addition, although we asked the participants about their voice discomfort and usages, we did not ask them about changes in their emotional stress, which might have affected their voices and influenced the results of this study (Helou, Rosen, Wang, & Verdolini Abbott, 2018).

Although the moderate reliability of RFF may have been related to actual changes in laryngeal function, it may also reflect the actual reliability of RFF and its current estimation process. The automated RFF algorithm used in the current study was developed as an alternative to the manual RFF estimation process, which is subjective and time-consuming. However, the current algorithm shows small differences in estimated RFF values compared to manual estimation (Lien et al., 2017). Improvements in automated RFF estimation to better detect offset and onset cycles may enhance the reliability of RFF for clinical use in the future. Another possibility to increase the reliability of RFF may be online monitoring of sound pressure level while producing RFF stimuli, since controlling intensity has been shown to increase the reliability of acoustic and aerodynamic measures (Lee et al., 1999). However, our previous work did not indicate that mean RFF values were significantly impacted by sound pressure level (Park & Stepp, 2018).

Conclusion

From recording individuals with healthy voices for five consecutive days, we found that RFF exhibited moderate test–retest reliability, which was slightly lower or comparable to commonly used acoustic and aerodynamic measures. We suspect that our finding of moderate reliability

Downloaded from: https://pubs.asha.org Boston University on 07/22/2019, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

may reflect, to some degree, actual changes in individuals' vocal function or tension, since RFF was affected by the time awake before the recording sessions. RFF was found to be more reliable than CAPE-V parameters, as assessed and performed by a voice-trained speech pathologist. In addition, RFF measured from soft voice recordings showed better reliability than those measured from comfortable voice. For future studies, sensitivity to change and minimal clinically important differences should be studied to further evaluate the appropriateness of RFF for clinical use.

Acknowledgments

This work was supported by Grant DC015570 from the National Institute on Deafness and Other Communication Disorders, awarded to Cara E. Stepp. The authors thank Talia Mittleman and Defne Abur for assistance with data recording, Daniel Buckley and Kimberly Dahl for performing Consensus Auditory–Perceptual Evaluation of Voice ratings, and Jessica Silfen, Lauren MacLellan, and Dante Cilento for help with data analysis.

References

American Speech-Language-Hearing Association. (2002). *Consensus Auditory–Perceptual Evaluation of Voice (CAPE-V): ASHA Special Interest Division 3, Voice and voice disorders*. Retrieved from https://www.asha.org/uploadedFiles/members/divs/D3CAPEVprocedures.pdf

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

Awan, S. N., Novaleski, C. K., & Yingling, J. R. (2013). Test–retest reliability for aerodynamic measures of voice. *Journal of Voice, 27*(6), 674–684.

Belafsky, P. C., Postma, G. N., & Koufman, J. A. (2002). Validity and reliability of the reflux symptom index (RSI). *Journal of Voice, 16*(2), 274–277.

Boersma, P., & Weenink, D. (2016). Praat: Doing phonetics by computer (Version 6.0.21) [Computer program]. Retrieved from http://www.praat.org/

Bough, I. D., Heuer, R. J., Sataloff, R. T., Hills, J. R., & Cater, J. R. (1996). Intrasubject variability of objective voice measures. *Journal of Voice, 10*(2), 166–174.

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.

Brockmann, M., Storck, C., Carding, P. N., & Drinnan, M. J. (2008). Voice loudness and gender effects on jitter and shimmer in healthy adults. *Journal of Speech, Language, and Hearing Research, 51*(5), 1152–1160.

Carding, P. N., Steen, I. N., Webb, A., MacKenzie, K., Deary, I. J., & Wilson, J. A. (2004). The reliability and sensitivity to change of acoustic measures of voice quality. *Clinical Otolaryngology, 29*(5), 538–544.

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

Chan, R. W., & Titze, I. R. (2006). Dependence of phonation threshold pressure on vocal tract acoustics and vocal fold tissue mechanics. *The Journal of the Acoustical Society of America, 119*(4), 2351–2362.

Chang, A., & Karnell, M. P. (2004). Perceived phonatory effort and phonation threshold pressure across a prolonged voice loading task: A study of vocal fatigue. *Journal of Voice, 18*(4), 454–466.

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.

Deliyski, D. D., Shaw, H. S., Evans, M. K., & Vesselinov, R. (2006). Regression tree approach to studying factors influencing acoustic voice analysis. *Folia Phoniatrica et Logopaedica, 58*(4), 274–288.

Desjardins, M., Halstead, L., Cooke, M., & Bonilha, H. S. (2017). A systematic review of voice therapy: What "effectiveness" really implies. *Journal of Voice, 31*(3), 392.e313–392.e332.

Eadie, T. L., & Baylor, C. R. (2006). The effect of perceptual training on inexperienced listeners' judgments of dysphonic voice. *Journal of Voice, 20*(4), 527–544.

Eadie, T. L., & Doyle, P. C. (2005). Classification of dysphonic voice: Acoustic and auditory–perceptual measures. *Journal of Voice, 19*(1), 1–14.

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

Enflo, L., Sundberg, J., Romedahl, C., & McAllister, A. (2013). Effects on vocal fold collision and phonation threshold pressure of resonance tube phonation with tube end in water. *Journal of Speech, Language, and Hearing Research, 56*(5), 1530–1538.

Eskenazi, L., Childers, D. G., & Hicks, D. M. (1990). Acoustic correlates of vocal quality. *Journal of Speech and Hearing Research, 33*(2), 298–306.

Espinoza, V. M., Zanartu, M., Van Stan, J. H., Mehta, D. D., & Hillman, R. E. (2017). Glottal aerodynamic measures in women with phonotraumatic and nonphonotraumatic vocal hyperfunction. *Journal of Speech, Language, and Hearing Research, 60*(8), 2159–2169.

Fairbanks, G. (1960). *Voice and articulation drillbook* (2nd ed.). New York, NY: Harper & Row.

Fryd, A. S., Van Stan, J. H., Hillman, R. E., & Mehta, D. D. (2016). Estimating subglottal pressure from neck-surface acceleration during normal voice production. *Journal of Speech, Language, and Hearing Research, 59*(6), 1335–1345.

Fujiki, R. B., Chapleau, A., Sundarrajan, A., McKenna, V., & Sivasankar, M. P. (2017). The interaction of surface hydration and vocal loading on voice measures. *Journal of Voice, 31*(2), 211–217.

Garrett, K. L., & Healey, E. C. (1987). An acoustic analysis of fluctuations in the voices of normal adult speakers across three times of day. *The Journal of the Acoustical Society of America, 82*(1), 58–62.

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. (2010). Madde (Version 3.0.0.2) [Computer program]: Tolvan data. Retrieved from http://www.tolvan.com/

Halpern, A. E., Spielman, J. L., Hunter, E. J., & Titze, I. R. (2009). The Inability to Produce Soft Voice (IPSV): A tool to detect vocal change in school-teachers. *Logopedics, Phoniatrics, Vocology, 34*(3), 117–127.

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.

Downloaded from: https://pubs.asha.org Boston University on 07/22/2019, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

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.

Helou, L. B., Rosen, C. A., Wang, W., & Verdolini Abbott, K. (2018). Intrinsic laryngeal muscle response to a public speech preparation stressor. *Journal of Speech, Language, and Hearing Research, 61*(7), 1525–1543.

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

Hillenbrand, J. (1988). Perception of aperiodicities in synthetically generated voices. *The Journal of the Acoustical Society of America, 83*(6), 2361–2371.

Hillman, R. E., Gress, C., Hargrave, J., Walsh, M., & Bunting, G. (1990). The efficacy of speech-language pathology intervention: Voice disorders. *Seminars in Speech and Language, 11,* 297–310.

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.

Holmberg, E. B., Perkell, J. S., & Hillman, R. E. (1984). Methods for using a noninvasive technique for estimating glottal functions from oral measurements. *The Journal of the Acoustical Society of America, 75*(S1). https://doi.org/10.1121/1.2021620

Hunter, E. (2011). General statistics of the NCVS Self-Administered Vocal Rating (SAVRa). *The National Center for Voice and Speech Online Memo*. Retrieved from http://www.ncvs.org/e-learning/tech/tech-memo-11.pdf

Jacobson, B. H., Johnson, A., Grywalski, C., Silbergleit, A., Jacobson, G., Benninger, M., & Newman, C. (1997). The voice handicap index (VHI): Development and validation. *American Journal of Speech-Language Pathology, 6*(3), 66–69.

Kagan, L. S., & Heaton, J. T. (2017). The effectiveness of low-level light therapy in attenuating vocal fatigue. *Journal of Voice, 31*(3), 384.e315–384.e323.

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.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163.

Lee, L., Stemple, J. C., & Kizer, M. (1999). Consistency of acoustic and aerodynamic measures of voice production over 28 days under various testing conditions. *Journal of Voice, 13*(4), 477–483.

Leong, K., Hawkshaw, M. J., Dentchev, D., Gupta, R., Lurie, D., & Sataloff, R. T. (2013). Reliability of objective voice measures of normal speaking voices. *Journal of Voice, 27*(2), 170–176.

Lien, Y. A., Gattuccio, C. I., & Stepp, C. E. (2014). Effects of phonetic context on relative fundamental frequency. *Journal of Speech, Language, and Hearing Research, 57*(4), 1259–1267.

Lien, Y. A., Heller Murray, E. S., Calabrese, C. R., Michener, C. M., Van Stan, J. H., Mehta, D. D., … Stepp, C. E. (2017). Validation of an algorithm for semi-automated estimation of voice relative fundamental frequency. *Annals of Otology, Rhinology & Laryngology, 126*(10), 712–716.

Lien, Y. A., 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.

Linder, R., Albers, A. E., Hess, M., Poppl, S. J., & Schonweiler, R. (2008). Artificial neural network-based classification to screen for dysphonia using psychoacoustic scaling of acoustic voice features. *Journal of Voice, 22*(2), 155–163.

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.

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.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30–46.

McKenna, V. S., Heller Murray, E. S., Lien, Y. S., & Stepp, C. E. (2016). The relationship between relative fundamental frequency and a kinematic estimate of laryngeal stiffness in healthy adults. *Journal of Speech, Language, and Hearing Research, 59*(6), 1283–1294.

Park, Y., & Stepp, C. E. (2018). The effects of stress type, vowel identity, baseline f_0, and loudness on the relative fundamental frequency of individuals with healthy voices. *Journal of Voice*. Epub ahead of print. Retrieved from https://sites.bu.edu/stepplab/files/2018/08/ParkSteppInPress.pdf

Patel, R. R., Awan, S. N., Barkmeier-Kraemer, J., Courey, M., Deliyski, D., Eadie, T., … 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.

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.

Solomon, N. P., & DiMattia, M. S. (2000). Effects of a vocally fatiguing task and systemic hydration on phonation threshold pressure. *Journal of Voice, 14*(3), 341–362.

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

Stemple, J. C., Stanley, J., & Lee, L. (1995). Objective measures of voice production in normal subjects following prolonged voice use. *Journal of Voice, 9*(2), 127–133.

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

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.

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.

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.

Stevens, K. N. (1977). Physics of laryngeal behavior and larynx modes. *Phonetica, 34*(4), 264–279.

Watson, B. C. (1998). Fundamental frequency during phonetically governed devoicing in normal young and aged speakers. *The Journal of the Acoustical Society of America, 103*(6), 3642–3647.

Downloaded from: https://pubs.asha.org Boston University on 07/22/2019, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions

Watts, C. R., Awan, S. N., & Maryn, Y. (2016). A comparison of cepstral peak prominence measures from two acoustic analysis programs. *Journal of Voice, 31*(3), 387.e1–387.e10.

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.

Welham, N. V., & Maclagan, M. A. (2003). Vocal fatigue: Current knowledge and future directions. *Journal of Voice, 17*(1), 21–30.

Zraick, R. I., Kempster, G. B., Connor, N. P., Thibeault, S., Klaben, B. K., Bursac, Z., ... Glaze, L. E. (2011). Establishing validity of the Consensus Auditory–Perceptual Evaluation of Voice (CAPE-V). *American Journal Speech-Language Pathollgy, 20*(1), 14–22.

Downloaded from: https://pubs.asha.org Boston University on 07/22/2019, Terms of Use: https://pubs.asha.org/pubs/rights_and_permissions