JSLHR

Research Article

Differences in Weeklong Ambulatory Vocal Behavior Between Female Patients With Phonotraumatic Lesions and Matched Controls

Jarrad H. Van Stan,[a,b,c] Daryush D. Mehta,[a,b,c] Andrew J. Ortiz,[a] James A. Burns,[a,b] Laura E. Toles,[a,c] Katherine L. Marks,[a,c] Mark Vangel,[a,b] Tiffiny Hron,[a,b] Steven Zeitels,[a,b] and Robert E. Hillman[a,b,c]

Purpose: Previous work using ambulatory voice recordings has shown no differences in average vocal behavior between patients with phonotraumatic vocal hyperfunction and matched controls. This study used larger groups to replicate these results and expanded the analysis to include distributional characteristics of ambulatory voice use and measures indicative of glottal closure.
Method: Subjects included 180 adult women: 90 diagnosed with vocal fold nodules or polyps and 90 age-, sex-, and occupation-matched controls with no history of voice disorders. Weeklong summary statistics (average, variability, skewness, kurtosis) of voice use were computed from neck-surface acceleration recorded using an ambulatory voice monitor. Voice measures included estimates of sound pressure level (SPL), fundamental frequency (f_o), cepstral peak prominence, and the difference between the first and second harmonic magnitudes (H1–H2).
Results: Statistical comparisons resulted in medium–large differences (Cohen's $d \geq 0.5$) between groups for SPL skewness, f_o variability, and H1–H2 variability. Two logistic regressions (theory-based and stepwise) found SPL skewness and H1–H2 variability to classify patients and controls based on their weekly voice data, with an area under the receiver operating characteristic curve of 0.85 and 0.82 on training and test sets, respectively.
Conclusion: Compared to controls, the weekly voice use of patients with phonotraumatic vocal hyperfunction reflected higher SPL tendencies (negatively skewed SPL) with more abrupt glottal closure (reduced H1–H2 variability, especially toward higher values). Further work could examine posttreatment data (e.g., after surgery and/or therapy) to determine the extent to which these differences are associated with the etiology and pathophysiology of phonotraumatic vocal fold lesions.

Phonotraumatic vocal hyperfunction (PVH) is a class of voice disorders characterized by clear signs of vocal fold tissue trauma on the medial/contact surfaces of the vocal folds (e.g., nodules, polyps; Mehta et al., 2015). The tissue trauma is believed to be caused and perpetuated by daily/habitual vocal behaviors that can include talking too loudly, using inappropriate pitch, talking too long without adequate rest/recovery, and/or employing inefficient phonation (e.g., generating higher-than-normal vocal fold collision forces to achieve a desired vocal intensity; Hillman et al., 1989; Karkos & McCormick, 2009; Kunduk & McWhorter, 2009; Leonard, 2009). The assumed relationship between daily vocal behaviors and PVH serves as the basis for current behavioral treatment approaches pursued as part of voice therapy. For example, vocal hygiene recommendations for patients with PVH include the introduction of voice rest periods, reductions in excessive voice use, and avoidance of talking over background noise or in rooms with excessive reverberation (Astolfi et al., 2015; Behlau & Oliveira, 2009; Bottalico et al., 2017; Holmberg et al., 2001; Roy et al., 2001, 2002). Unfortunately, assumptions about the role of daily voice use in the etiology of PVH have still not been adequately verified or objectively delineated, which continues to hamper the effective prevention and evidence-based management of this common voice disorder.

Ambulatory voice monitoring technology has the potential to examine the role of voice use in PVH by providing

[a]Massachusetts General Hospital, Boston
[b]Harvard Medical School, Boston, MA
[c]MGH Institute of Health Professions, Boston, MA
Correspondence to Jarrad H. Van Stan: jvanstan@mgh.harvard.edu

Editor-in-Chief: Bharath Chandrasekaran
Editor: Jack J. Jiang
Received July 12, 2019
Revision received September 18, 2019
Accepted October 14, 2019
https://doi.org/10.1044/2019_JSLHR-19-00065

Disclosure: *Robert E. Hillman and Steven Zeitels have a financial interest in InnoVoyce LLC, a company focused on developing and commercializing technologies for the prevention, diagnosis, and treatment of voice-related disorders. Hillman's and Zeitels' interests were reviewed and are managed by Massachusetts General Hospital and Partners HealthCare in accordance with their conflict of interest policies.*

the means to objectively characterize habitual vocal behavior during activities of daily living (Carullo et al., 2013; Cheyne et al., 2003; Popolo et al., 2005; Searl & Dietsch, 2014; Szabo et al., 2001). Such devices typically employ a neck-placed sensor—often a miniature accelerometer (ACC)—to sense neck-skin vibration to unobtrusively monitor phonation (Van Stan et al., 2014). To date, these devices have been mostly used to characterize the vocal demands of speakers with healthy vocal status in occupations that have a higher-than-normal risk of developing a voice disorder (e.g., teachers, singers, telemarketers; Calosso et al., 2017; Carroll et al., 2006; Hunter & Titze, 2009, 2010; Lindstrom et al., 2011; Morrow & Connor, 2011; Puglisi et al., 2017; Södersten et al., 2005). Because a higher risk of developing a voice disorder (particularly related to phonotrauma) is hypothetically associated with speaking too loudly, at an inappropriate pitch, and/or too much with inadequate vocal rest, ambulatory voice monitors have traditionally measured subjects' vocal intensity, fundamental frequency (f_o), and amount of voice use (i.e., vocal dose) as overall averages, standard deviations, and/or total accumulations (Bottalico & Astolfi, 2012; Bottalico et al., 2018; Carroll et al., 2006; Carullo et al., 2015; Ghassemi et al., 2014; Hillman et al., 2006; Hunter & Titze, 2010; Mehta et al., 2015; Titze & Hunter, 2015; Titze et al., 2007; Van Stan et al., 2015). Vocal dose measures attempt to indirectly estimate the exposure of vocal fold tissue to mechanical stress during phonation. Frequently used dose measures include the estimation of accumulated phonation time (time dose), the number of true vocal fold oscillatory cycles (cycle dose), and the total distance traveled by the vocal folds (distance dose) that combines intensity, f_o, and phonation time (Švec et al., 2003; Titze et al., 2003). The general concept of vocal dose is based on occupation safety standards for vibration exposure to various body structures (e.g., noise exposure and hearing loss, jackhammer use, and musculoskeletal disorders of the upper extremities).

To date, only a few studies have used ambulatory voice monitoring technology to investigate differences in average daily vocal behavior between patients with PVH and matched controls (Cortés et al., 2018; Ghassemi et al., 2014; Maffei et al., 2016; Masuda et al., 1993; Mehta et al., 2015; Nacci et al., 2013; Szabo Portela et al., 2018; Van Stan et al., 2015). Contrary to clinical intuition about the vocal behavior of patients with PVH, none of the studies identified significant differences in average vocal intensity, f_o, and vocal doses between the two groups. Mehta et al. (2015) also reported no difference between patients with PVH and matched controls for average measures of cepstral peak prominence (CPP) extracted from the neck ACC signal. It has since been verified that such ACC-based measures of CPP are highly correlated with the measures of CPP from the acoustic (microphone) signal (Mehta et al., 2016)—and acoustic CPP is recommended for clinical use to quantify the level of periodic energy in the acoustic voice signal (Patel et al., 2018). This recommendation is supported by evidence that CPP is highly correlated with clinician auditory–perceptual ratings of overall dysphonia (Awan et al., 2010). Thus, the lack of a significant difference in CPP between patients with PVH and controls also appears to run counter to the clinical expectation that patients with vocal fold lesions are more dysphonic than healthy speakers.

The only consistent significant difference in a weekly average voice statistic has been f_o variability (patients voiced with less variability, especially less variance toward higher frequencies; Mehta et al., 2015; Van Stan et al., 2015). Other analysis approaches have quantified trends over time with inconsistent results—patients' decreased mean vocal intensity and f_o over time (Nacci et al., 2013) or increased both over time (Ghassemi et al., 2014)—and investigations into the relationship between patient-reported vocal status improvement/decrement and objective ambulatory measures have found no consistent, unidimensional associations across patients (Maffei et al., 2016). However, better-than-chance classification of patients with PVH and matched controls has been done using extreme distributional characteristics (e.g., 5th and 95th percentiles) and advanced machine learning algorithms (Ghassemi et al., 2014). Therefore, it may be possible that "average" behavior differences could be represented in more subtle characteristics of weekly distributions, that is, higher order moments such as skewness or kurtosis. For example, if the patient is talking more often in a slightly louder part of their range than a matched control (not constantly talking louder than the control), the louder behavior will be represented by a change in skew but not in the mean, median, or mode of the distribution. Alternatively, if patients with PVH talk with less extreme variability (not average variability), distributions might be better represented by kurtosis than by the standard deviation. In a similar vein, for a measure like CPP, if patients are inconsistently more dysphonic and only produce episodes of dysphonia, then a difference in overall voice quality will not be represented by the mean but by higher order estimates of the distribution such as skewness or kurtosis. For example, one study, which used sustained vowels recorded in the laboratory, achieved better classification between a small sample of controls ($n = 35$) and patients with a variety of voice disorders ($n = 41$) with CPP 5th percentile than CPP mean (Castellana et al., 2018).

The lack of consistent differences between patients with PVH and controls in traditional measures of vocal intensity, f_o, CPP, and vocal dose could result from the patients compensating to maintain functional values of these parameters in the presence of phonotraumatic lesions (i.e., maladaptive compensation). Multiple laboratory studies have shown that patients with PVH produce phonation with higher potential for vocal fold trauma/contact than matched controls (e.g., higher subglottal pressure, maximum flow declination rate, and/or unsteady flow) while maintaining normal average values for sound pressure level (SPL) and f_o (Espinoza et al., 2017; Hillman et al., 1989; Holmberg et al., 2003). Therefore, it would be desirable to investigate additional measures that can also be extracted from the ACC signal (neck-placed ambulatory phonation sensor) and can provide additional insights into underlying

phonatory mechanisms. One such measure is the difference (in dB) between the levels of the first and second harmonics (H1–H2).

H1–H2 is a low-bandwidth measure of spectral tilt that is commonly used as an acoustic-based estimate of vocal fold closure during phonation (Klatt & Klatt, 1990; Stevens, 1998). Changes in H1–H2 have been correlated to the abruptness of glottal closure (i.e., skewness of the glottal airflow pulse), open quotient, and the dimension of breathy-to-strained voice quality (Henrich et al., 2001; Hillenbrand et al., 1994; Klatt & Klatt, 1990; Lowell et al., 2012; Swerts & Veldhuis, 2001; Zhang, 2016). Larger differences between the two harmonics (higher H1–H2) are associated with a glottal vibratory pattern exhibiting less abrupt/reduced vocal fold closure and breathier voice quality; smaller differences (lower H1–H2) are associated with more abrupt/increased vocal fold closure and more strained voice quality. Furthermore, H1–H2 has great potential to differentiate patients with PVH from matched controls, as Cortes et al. (2018) recently showed better-than-chance classification between these two groups where H1–H2 kurtosis was the largest contributor; patients voiced with much less extreme variability (higher H1–H2 kurtosis) than matched controls. Finally, of relevance to this study, it has recently been shown that H1–H2 measures extracted from the raw ACC signal correlate highly with H1–H2 measures from the inverse-filtered oral airflow signal (r = .72; Mehta et al., 2019). The high correlation offers the possibility of being able to interpret ACC-based measures of H1–H2 as an indirect indicator of glottal closure.

The purpose of this study was twofold: (a) use larger groups of subjects to verify previous results that average measures of SPL, f_o, CPP, vocal dose, and phonatory/nonphonatory segments acquired from daily life were not significantly different between patients with PVH and matched controls (except f_o variability) and (b) determine if there are significant differences in daily vocal behavior using a physiologically salient measure of vocal function (H1–H2) and higher order distributional characteristics of SPL, f_o, CPP, and H1–H2. Weeklong ambulatory phonation data were acquired using a smartphone-based ambulatory voice monitor (using an ACC as the phonation sensor; Mehta et al., 2012) in groups of patients with PVH and age-, gender-, and occupation-matched controls that were large enough to provide adequate power for robust statistical testing of even weak/small differences between groups. All data were collected as part of a larger, ongoing project aimed at attaining a better understanding of the etiology and pathophysiology of hyperfunctional voice disorders. The governing institutional review board approved all experimental aspects related to the use of human subjects for this study.

Method

Participants

One hundred eighty total female subjects were consented for participation in this study. Ninety female patients with vocal fold nodules or polyps were recruited through sequential convenience sampling. Only female participants were selected to be in this study to provide a homogenous sample of a group that has a significantly higher incidence of phonotraumatic vocal fold lesions (Goldman et al., 1996; Herrington-Hall et al., 1988). Diagnoses were based on a comprehensive team evaluation (laryngologist and speech-language pathologist) at the Center for Laryngeal Surgery and Voice Rehabilitation at Massachusetts General Hospital (MGH Voice Center) that included (a) the collection of a complete case history, (b) endoscopic imaging of the larynx, (c) completion of the Voice-Related Quality of Life (V-RQOL) questionnaire (Hogikyan & Sethuraman, 1999), (d) an auditory–perceptual evaluation using the Consensus Auditory–Perceptual Evaluation of Voice (CAPE-V; (Kempster et al., 2009), and (e) aerodynamic and acoustic assessments of vocal function. A control subject with no history of voice disorders was matched to each patient according to approximate age (± 5 years), sex, and occupation. The normal vocal status of all control participants was verified via interview and a laryngeal stroboscopic examination. During the interview, the matched-control candidates were specifically asked if they had any voice difficulties that affected their daily life, and a speech-language pathologist evaluated the auditory–perceptual quality of their voices. If the matched-control candidate indicated voice difficulties or demonstrated a nonnormal voice quality, they were excluded from study enrollment and did not undergo a laryngeal stroboscopic examination.

Of the 90 patients, 79 were diagnosed with bilateral vocal fold nodules, eight were diagnosed with a unilateral vocal fold polyp, two were diagnosed with a unilateral vocal fold polyp and reactive vocal fold nodule, and one was diagnosed with bilateral vocal fold nodules and a left vocal fold polyp. All participants were engaged in occupations considered to be at a higher-than-normal risk for developing a voice disorder (Verdolini & Ramig, 2001). The majority of patient–control pairings were professional, amateur, or student singers (67 pairs); all patient singers were matched with control subjects who were in the same musical genre (classical or nonclassical) to account for any genre-specific vocal behaviors. The other occupations included administrator (three pairs), teacher (two pairs), psychologist (two pairs), talent recruiter (two pairs), registered nurse (one pair), retiree (one pair), media relations (one pair), marketer (one pair), and consultant (one pair). The average (standard deviation) age of participants within each group was approximately 26 (10) years.

Table 1 reports subscale scores for the self-reported V-RQOL and clinician-judged CAPE-V ratings for the participants in the patient group. V-RQOL scores are normalized ordinal ratings that lie between 0 and 100, with higher scores indicating a higher quality of life. CAPE-V scores are visual analog scale ratings that range from 0 to 100, with zero indicating normality and 100 indicating extremely severe abnormality of a particular voice quality characteristic. Scores on both perceptual scales indicated that most participants exhibited mild-to-moderate

Table 1. Patients' self-reported quality of life impact due to their voice disorder using the Voice-Related Quality of Life (V-RQOL) subscales and the perceived qualities of their voice as judged by a speech-language pathologist using the Consensus Auditory–Perceptual Evaluation of Voice (CAPE-V) form.

Measure	*M* ± *SD*
V-RQOL	
Social–emotional	73.5 ± 22.0
Physical functioning	72.2 ± 19.5
Total score	72.5 ± 17.8
CAPE-V	
Overall severity	26.8 ± 14.8
Roughness	18.5 ± 14.6
Breathiness	14.1 ± 12.6
Strain	19.5 ± 12.9
Pitch	6.8 ± 10.4
Loudness	3.9 ± 8.6

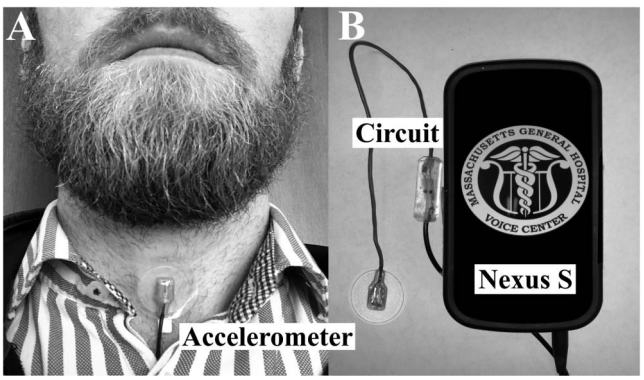
Note. Mean and standard deviation reported (n = 90).

dysphonia, with only a few falling on the very severe end of the scales.

Data Collection

The Voice Health Monitor (VHM; Mehta et al., 2012) was used to collect ambulatory voice data on all subjects in the study. As shown in Figure 1, the VHM employs a miniature ACC (Model BU-27135, Knowles Electronics) attached via double-sided medical grade tape to the anterior neck (below the larynx and above the sternal notch) to sense phonation. The sensor is connected to a custom smartphone application as the data acquisition platform, and the system records the unprocessed acceleration signal at 11,025-Hz sampling rate, 16-bit quantization, and 80-dB dynamic range to obtain frequency content of neck-surface vibrations up to 5 kHz. The VHM application provides a user-friendly interface for starting/stopping recording, daily sensor calibration, periodic alert capabilities that include system checks (Mehta et al., 2012), and vocal status questions (e.g., asking users about their level of vocal fatigue; Van Stan et al., 2017).

Figure 1. Illustration of the accelerometer-based ambulatory voice monitor: (A) wired accelerometer mounted on a silicone pad affixed to the anterior neck surface midway between the thyroid prominence and the suprasternal notch and (B) smartphone, accelerometer sensor, and interface cable with circuit encased in epoxy.



Participants in the patient group were monitored for 1 week (7 days) before any surgical and/or therapeutic intervention. Each control participant was monitored for 1 full week. Each morning, the VHM application led the participants through a daily process to calibrate the ACC signal level to acoustic SPL recorded by a handheld microphone (H1 Handy Recorder, Zoom Corporation) positioned 15 cm from the lips (Švec et al., 2005; Van Stan et al., 2015). For the acoustic SPL calibration, the participant is asked to glide from soft to loud on an /a/ and is trained to perform the loudness glide during their initial study appointment by study staff. To also improve the quality of the loudness glide, three glides are elicited from the subjects every morning and the best glide (largest intensity range and most linear mapping between the neck skin and acoustic signal) is used. The most detailed description of the acoustic SPL calibration is included in a previous publication (Mehta et al., 2012). During the calibration procedure, participants take a picture of their neck to document the day-to-day placement of ACC. Participants were also taught to contact study staff if the sensor fell off their neck or loosened throughout the day. If the ACC sensor was misplaced (as evidenced by the daily photos) or the participants reported issues with the sensor coming off, those days of data were not included in the analysis. Of note, these types of sensor issues occurred very rarely.

Data Analysis

Before processing the hours-long neck-skin acceleration recordings, SPL calibration factors (multiplier and offset) are computed to transform the neck-skin acceleration amplitude into an estimate of acoustic SPL. Specifically, a linear regression is computed for everyday of monitoring by time-aligning the neck-skin acceleration signal and acoustic SPL signal for each loudness glide recorded during the morning calibration procedure. Each signal is processed using nonoverlapping 50-ms analysis windows. Once the SPL calibration factors have been computed, they are used to process the ambulatory recordings.

The hours-long neck-skin acceleration recordings were divided into nonoverlapping frames of 50 ms in duration. As was done in previous studies (Mehta et al., 2015; Van Stan et al., 2015), each frame was considered voiced if it passed the following thresholds: (a) SLP was greater than 45 dB SPL at 15 cm, (b) the first nonzero-lag peak in the normalized autocorrelation exceeded a threshold of 0.6, (c) f_o (reciprocal of the time lag of the first nonzero autocorrelation peak) was between 70 and 1000 Hz, and (d) the ratio of low- to high-frequency energy exceeded 20 dB. These criteria were needed to eliminate several types of nonphonatory activity such as tapping or rubbing on the

sensor, extremely high levels of environmental noise (e.g., rock concert), and electrical interference/artifacts.

CPP and H1–H2 were two additional features calculated on each analysis frame. To calculate CPP, each 50-ms frame underwent two discrete Fourier transforms that were computed in succession with a logarithmic transformation between them. A regression line was then computed over quefrencies greater than 2 ms (corresponding to a quefrency range minimally affected by subglottal resonances). Finally, the CPP for each frame was defined as the difference, in dB, between the magnitude of the highest peak and the baseline regression level in the power cepstrum. The peak search was limited to quefrencies between 2.5 and 12 ms, corresponding to frequencies of 417 and 83 Hz, respectively. To calculate H1–H2, each 50-ms frame underwent one discrete Fourier transform. The H1–H2 for each frame was defined as the difference, in dB, between the amplitudes of the first and second harmonics in the frequency spectrum.

Three cumulative vocal dose measures represented each participant's average voice use: phonation time, cycle dose, and distance dose. Phonation time was the total duration (sum) of each 50-ms frame classified as "voiced" during the total monitoring time. Cycle dose estimated the total number of vocal fold oscillations during the monitored time by summing all voiced frames according to f_o (higher f_o would be represented as more vocal fold oscillations). Finally, distance dose estimated the total distance traveled (in meters) by the vocal folds by multiplying cycle dose with estimates of vibratory amplitude based on SPL (Švec et al., 2003).

Lastly, grounded in previous approaches, temporal measures of vocal load and recovery time were categorized according to the occurrences and durations of contiguous voiced and nonvoiced segments (Titze et al., 2007). Voiced and nonvoiced segment durations were binned into logarithmically spaced ranges from 0.100–0.316 to 3,160–10,000 s, where successively longer duration segments represented successively higher level speech segmentals (phoneme level, syllable level, word level, etc., for voiced segments; voiceless consonants, pauses between phrases, etc., for nonvoiced segments) up to the longest duration sung passages and silence periods. These data yielded two types of histograms: (a) "occurrence" histograms of the normalized (per-hour) counts of all contiguous voiced and nonvoiced segments within each duration bin and (b) "accumulation" histograms of the total duration (normalized per hour) of all contiguous voiced and nonvoiced segments within each duration bin. A count of phonatory onsets per hour was derived from the total number of voiced segments divided by the total number of hours monitored.

Statistical Analysis

Within-subject univariate summary statistics characterized the distributions of weeklong SPL, f_o, CPP, and H1–H2 time series of lengths ranging from 200,000 to over 1,000,000 voiced frames, depending upon how much subjects phonated during their respective weeks. Statistics computed were mean (SPL, CPP, and H1–H2), mode (f_o only), standard deviation, minimum (5th percentile), maximum (95th percentile), range (middle 90%), skewness, and kurtosis. In the data presented here, SPL, CPP, and H1–H2 distributions tended to be normal (similar mean, median, and mode), and f_o distributions were often skewed toward lower f_o values with a long, thin tail toward higher f_o values. The f_o mode was computed from histograms containing 30 equally spaced bins.

Vocal dose measures were computed as both total accumulated values over the entire monitored time for each individual and normalized values to account for differences in total time monitored by each subject. From the occurrence and accumulation histograms for phonatory/nonphonatory segments, per-hour counts and durations of voiced and nonvoiced segments within each duration bin were recorded for each participant.

To take full advantage of the matched patient–control paradigm ($n = 90$ pairs), paired t tests (parametric data) and Wilcoxon signed-ranks tests (nonparametric data) were used to assess differences between the summary statistics of weekly voice use. A Kolmogorov–Smirnov (KS) test was used to assess the normalcy (parametric distribution) of each distribution of paired differences (patient minus control). When the KS test was significant ($p < .05$), a Wilcoxon signed-ranks test evaluated the distribution of paired differences against the null hypothesis of zero (i.e., "no difference"). When the KS test was not significant, a paired t test evaluated the distribution of paired differences against the null hypothesis of zero. Due to the large number of tests, the alpha level of significance was adjusted using a Bonferroni approach ($\alpha = .0014$ and .0016 for voiced features and phonatory/nonphonatory segments, respectively). When statistical significance was found, the difference was characterized by a Cohen's d effect size calculation. For example, the difference between the two groups' means divided by their pooled standard deviation. Cohen's d provided a standardized method to interpret the degree of differences between the two groups (small when $\leq$ 0.19, small to medium when 0.20–0.49, medium to large when 0.50–0.79, and large when $\geq$ 0.80; Cohen, 1988).

A partially theory-driven logistic regression model was trained and tested using the most predictive features (the stepwise logistic regression only contained features with medium-to-large Cohen's d effect sizes). Since only the features with medium-to-large effect sizes were used in the partially theory-driven model, it is possible to train a better model using all statistically significant features (a completely data-driven approach). For example, perhaps, a combination of one strong predictor and one weak predictor (e.g., SPL skew and percent phonation, respectively) would improve model performance? Therefore, a fully data-driven, stepwise logistic regression was trained that used all significant features (regardless of effect size). For both stepwise logistic regressions, a forward, conditional approach was chosen to minimize the total number of features and feature redundancy (i.e., minimal correlation between final variables). The models was first trained on half of the data

set (45 patient–control pairs) and then tested on the second half of the data (a held-out set of 45 patient–control pairs). The training and test sets were equally balanced according to the number of singers and nonsingers (33 and 34 pairs, respectively) and voice quality severity according to the treating clinician's CAPE-V rating of overall dysphonia. The ratings of overall dysphonia were (mean, standard deviation, and range) 25.8, 14.7, and 0–59 for the training set and 26.2, 15.3, and 0–69 for the test set. The two logistic regression models were considered statistically similar if the 95% confidence intervals (CIs) for their area under the receiver operating characteristic curves (AUCs) overlapped.

Results

Most subjects wore the monitoring system for more than 80 hr during the 7 days. Ten features produced distributions of paired differences that were nonnormal: monitored time, f_o mode, f_o 5th percentile, f_o kurtosis, cumulative cycle dose, and phonatory segments of 1–3.16 and 3.16–10 s (both occurrences and accumulations). Table 2 displays all summary statistics for voiced features (SPL, f_o, CPP, H1–H2, and vocal dose measures) that were compared between the patient and control groups. Ten measures were significantly different between the two groups ($p < .0014$): SPL skew, f_o variability (standard deviation, 95th percentile, range, and kurtosis), H1–H2 variability (standard deviation, 95th percentile, range, and kurtosis), and percent phonation time. Specifically, patients exhibited significantly more negative SPL skew ($d = 0.56$), lower overall variability and less variation toward higher f_o values ($|d| = 0.43$–0.67), lower overall variability and less variation toward higher H1–H2 values ($|d| = 0.74$–0.88), and higher percent phonation time ($|d| = 0.35$) compared to their matched controls.

Table 3 displays all features compared between patients and their matched controls from the phonatory and nonphonatory segment analysis. Fourteen measures were significantly different between the two groups ($p < .0018$): phonatory onsets per hour, phonatory and nonphonatory segments in a 0.1- to 0.316-s bin (both occurrences and

Table 2. Group-based mean (standard deviation) for weekly summary statistics of ambulatory estimates of sound pressure level (SPL), fundamental frequency (f_o), cepstral peak prominence (CPP), and H1–H2 measures collected from the patient and matched-control groups ($n = 90$ pairs).

Voice use summary statistic	Patient group	Control group	Cohen's *d*
Monitored duration (hr:min)	80:58 (18:32)	87:56 (14:48)	
SPL (dB SPL re 15 cm)			
M	85.8 (4.6)	84.5 (5.1)	
SD	11.5 (2.2)	12.1 (2.4)	
5th percentile	65.8 (5.4)	64.2 (6.3)	
95th percentile	104.0 (6.8)	104.6 (6.9)	
Range	38.1 (7.7)	40.4 (8.4)	
Skewness	–0.249 (0.272)	–0.033 (0.298)	0.56
Kurtosis	3.23 (0.44)	3.04 (0.38)	
f_o (Hz)			
Mode	196.1 (23.2)	199.4 (19.1)	
SD	73.5 (15.7)	86.7 (21.3)	0.66
5th percentile	165.3 (18.2)	168.6 (15.6)	
95th percentile	383.8 (58.8)	430.8 (78.1)	0.65
Range	218.5 (52.6)	262.1 (70.4)	0.67
Skewness	1.958 (0.560)	1.766 (0.505)	
Kurtosis	10.01 (5.25)	7.74 (3.16)	–0.43
CPP (dB)			
M	23.1 (1.2)	22.7 (1.1)	
SD	4.4 (0.3)	4.4 (0.3)	
5th percentile	15.2 (0.6)	15.0 (0.6)	
95th percentile	29.6 (1.3)	29.4 (1.3)	
Range	14.4 (0.9)	14.4 (1.1)	
Skewness	–0.281 (0.190)	–0.224 (0.189)	
Kurtosis	2.44 (0.18)	2.39 (0.16)	
H1–H2 (dB)			
M	4.4 (1.7)	5.1 (2.0)	
SD	6.1 (0.8)	7.0 (0.8)	0.88
5th percentile	–3.9 (2.1)	–4.3 (2.1)	
95th percentile	15.9 (2.5)	18.6 (2.5)	0.81
Range	19.8 (2.9)	22.9 (2.7)	0.86
Skewness	0.737 (0.315)	0.699 (0.254)	
Kurtosis	4.36 (0.86)	3.61 (0.61)	–0.74

Note. Comparisons reaching statistical significance ($p < .0014$) have Cohen's *d* effect sizes listed. Directionality of effect sizes is derived from the pairwise comparison of each summary statistic for control values minus their matched patient values.

Table 3. Group-based values of mean (standard deviation) of occurrence and accumulation of phonatory and nonphonatory segment duration bins for patients and controls (n = 90 pairs).

Voice use summary statistic	Patient group	Control group	Cohen's *d*
Phonatory segments			
Onsets (per hour)	1240 (375)	1073 (361)	–0.38
Occurrences (per hour)			
0.1–0.316 s	903 (282)	788 (272)	–0.35
0.316–1 s	310 (95)	251 (96)	–0.48
1–3.16 s	27 (16)	26 (20)	
3.16–10 s	1.9 (1.6)	2.6 (2.7)	
Accumulation (seconds per hour)			
0.1–0.316 s	154 (47)	133 (46)	–0.39
0.316–1 s	155 (48)	125 (49)	–0.47
1–3.16 s	40 (25)	39 (31)	
3.16–10 s	8.1 (7.0)	11.6 (12.8)	
Nonphonatory segments			
Occurrences (per hour)			
0.1–0.316 s	492 (169)	422 (166)	–0.36
0.316–1 s	186 (62)	159 (58)	–0.38
1–3.16 s	142 (38)	120 (37)	–0.42
3.16–10 s	77 (18)	66 (19)	–0.41
10–31.6 s	29 (6)	26 (7)	
31.6–100 s	9 (2)	9 (2)	
100–316 s	3.0 (0.7)	3.0 (0.8)	
316–1,000 s	0.89 (0.33)	1.00 (0.35)	
1,000–3,160 s	0.19 (0.13)	0.25 (0.15)	
Accumulation (seconds per hour)			
0.1–0.316 s	77 (26)	67 (25)	–0.35
0.316–1 s	106 (35)	90 (32)	–0.39
1–3.16 s	253 (67)	214 (66)	–0.42
3.16–10 s	423 (102)	368 (104)	–0.40
10–31.6 s	480 (116)	444 (133)	
31.6–100 s	478 (112)	474 (154)	
100–316 s	511 (118)	515 (141)	
316–1,000 s	458 (175)	531 (200)	
1,000–3,160 s	293 (200)	395 (252)	0.36

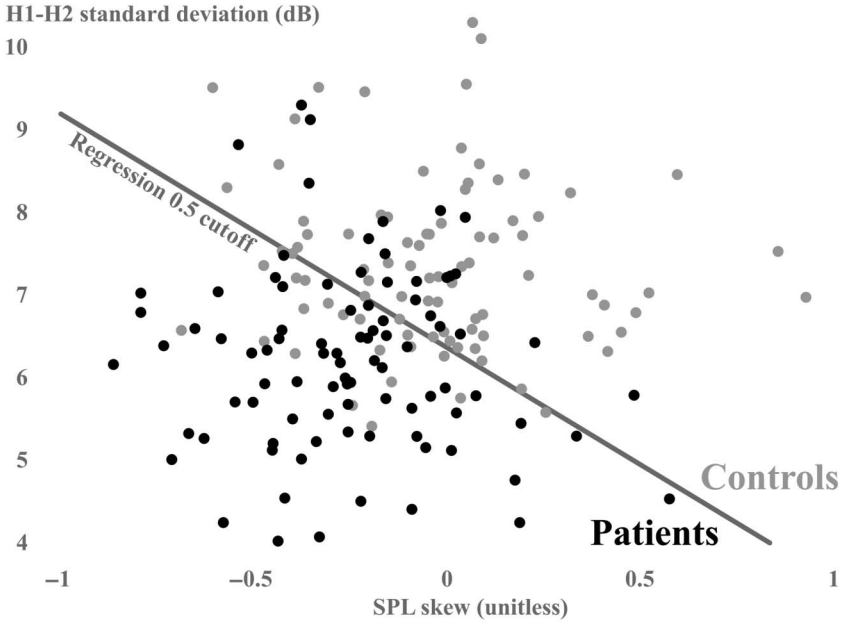
Note. Comparisons reaching statistical significance ($p < .0018$) have Cohen's *d* effect sizes. Directionality of effect sizes is derived from the pairwise comparison of each summary statistic for control values minus their matched patient values.

accumulation per hour), phonatory and nonphonatory segments in a 0.316- to 1-s bin (both occurrences and accumulation per hour), nonphonatory segments in a 1- to 3.16-s bin (both occurrences and accumulation per hour), nonphonatory segments in a 3.16- to 10-s bin (both occurrences and accumulation per hour), and accumulation of nonphonatory segments in a 1,000- to 3,160-s bin. Specifically, patients exhibited significantly more phonatory onsets per hour (d = 0.38), more short phonatory (< 1 s; d = 0.36–0.48) and nonphonatory (< 10 s) segments (d = 0.35–0.42), and less long nonphonatory segments (d = 0.36) compared to their matched controls.

A partially theory-driven logistic regression used only features with medium-to-large effect sizes ($d \geq 0.5$): SPL skew, f_o standard deviation, and H1–H2 standard deviation. Of note, standard deviation was used to represent the variability of f_o and H1–H2 because it is a simpler statistic to interpret than kurtosis, requires less data than kurtosis and extreme (5th and 95th) percentiles, and was highly correlated to all other variability metrics (Pearson r = .63–.99). Only two features were significant contributors to the model based on the training data of 45 patient–control pairs: SPL skew (b weight = −3.178, odds ratio [OR] = 0.042, p = .002) and H1–H2 standard deviation (b weight = −1.516, OR = 0.219, $p < .001$). The resulting overall classification for the training set was 74.4%, true positives = 36 subjects, true negatives = 31 subjects, false positives = 14 subjects, false negatives = nine subjects, and AUC = 0.846 (95% CI [0.768, 0.924]). The resulting overall classification for the test set was 76.7%, true positives = 33 subjects, true negatives = 36 subjects, false positives = nine subjects, false negatives = 12 subjects, and AUC = 0.823 (95% CI [0.736, 0.910]). The Pearson correlation coefficient between the two final variables was nonsignificant (r = .103). Figure 2 plots H1–H2 standard deviation against SPL skew to illustrate the performance of the two-variable model on classification of each subject based on weekly data (combined training and test set).

A data-driven stepwise logistic regression selected two measures from the 24 total significant measures to classify the training data: SPL skew (b weight = −3.321, OR = 0.040, p = .003) and H1–H2 range (b weight = −0.428,

Figure 2. Scatter plots of H1–H2 standard deviation on the *y*-axis and sound pressure level (SPL) skew on the *x*-axis (patients with vocal hyperfunction: black; matched controls: gray). Each dot represents a single patient's weekly distribution. The logistic regression cutoff is represented as a gray diagonal line.



$OR = 0.652$, $p < .001$). The resulting overall classification for the training set was 76.7% (true positives = 33, true negatives = 36, false positives = 9, false negatives = 12, AUC = 0.821), and that for the test set was 76.7% (true positives = 36, true negatives = 33, false positives = 12, false negatives = 9, AUC = 0.843). Pearson correlation coefficient between the two final variables was nonsignificant ($r = .096$). Based on the overlap in the AUC 95% CIs between the partially theory-driven and data-driven logistic regressions, the models were not significantly different in performance. Also of note, the different H1–H2 variability metrics (standard deviation and range) appear to be redundant with one another, as they are very highly correlated ($r = .993$).

Discussion

One purpose of this study was to use larger groups of subjects to verify previous results that average measures of SPL, f_o, CPP, and vocal dose acquired from daily life were not significantly different between patients with PVH and matched controls (except f_o variability). Reduced f_o variability, especially toward high frequencies, in patients was replicated in both statistical significance and the strength of the difference as measured by Cohen's d (Mehta et al., 2015; Van Stan et al., 2015). Lower f_o variability is likely related to the observation that patients with phonotraumatic lesions have a decreased ability to reach higher frequencies due to reduced pliability of the vocal fold lamina propria (Zeitels et al., 2002).

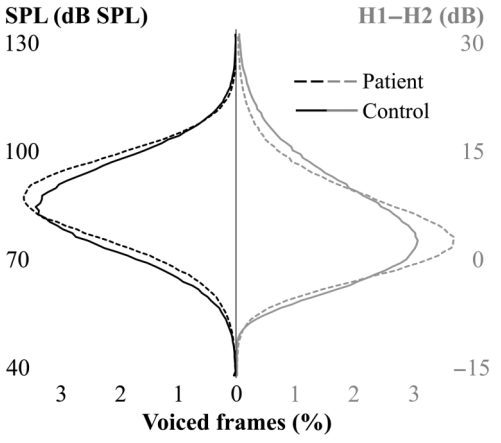
The replication of nonsignificant differences for average SPL, f_o, and CPP values verifies previous findings that these average measures associated with loudness, pitch, and dysphonia do not consistently differentiate between patients with phonotraumatic vocal fold lesions and matched healthy controls in terms of daily voice use. On the one hand, this appears to contradict the classic view that phonotrauma is typically associated with excessive loudness, inappropriate pitch, and obvious levels of dysphonia. However, in terms of underlying pathophysiology, the lack of difference also suggests that patients with phonotrauma are compensating for the presence of vocal fold lesions to maintain functional/acceptable levels of vocal loudness, pitch, and quality. Several laboratory and modeling studies have demonstrated that patients with phonotraumatic lesions seem to employ phonatory adjustments that maintain vocal SPL at the expense of increased potential for vocal fold trauma (e.g., elevated airflow and subglottal pressure metrics; Espinoza et al., 2017; Hillman et al., 1989; Zañartu et al., 2014). Thus, a lack of difference between patients and matched controls in SPL, f_o, and CPP does not necessarily mean that the underlying mechanisms for achieving these outputs remain equivalent in both groups.

Measures that traditionally attempt to characterize vocal dose and recovery (e.g., percent phonation, cycle dose, distance dose, and phonatory/nonphonatory segments) were either statistically indistinguishable between patients and controls or exhibited small effect sizes. For example, although percent phonation time reached a level of statistical significance, the effect size was small to medium ($d = 0.35$) and patients (10.0%) were approximately 1 percentage point

higher than that of their matched controls (8.6%). The significant differences reported for multiple phonatory and nonphonatory segments also demonstrated small-to-medium effect sizes ($|d|$ = 0.35–0.48). Lastly, despite statistically significant differences, none of the vocal dose or segmental features contributed to the logistic regression model. In general, it appears that the two-variable model characterizing vocal behavior (including SPL skew and H1–H2 standard deviation) provided much stronger discrimination between patients and controls than how much voicing or voice rest occurred. However, it would be premature to abandon the vocal dose and segmental measures since the weak differences that were observed may indicate that the measures could still be useful in characterizing important vocal behaviors in some individuals or subgroups of individuals (e.g., pre- vs. posttreatment). In fact, one article using segmental measures found that teachers with voice disorders spoke with significantly higher amounts of voicing across multiple segmental bins (Bottalico et al., 2017). Finally, one reason for the lack of differences (or weak differences) could be that these simple dose estimates do not include the amount of vocal fold contact or collision during voicing, which is the hypothesized causative and/or associative feature of phonotraumatic lesions. Thus, these results call for future work to develop vocal doses that incorporate key etiologic factors of phonotrauma, such as vocal fold collision or vocal fold closure parameters.

The second purpose of this study was to determine if there were significant differences in daily vocal behavior between patients with PVH and matched controls in traditional lower order distributional statistics of H1–H2 and higher order distributional characteristics of SPL, f_o, CPP, and H1–H2. SPL skew differences between patients and controls resulted in a medium-to-large effect size (d = 0.56), and the strongest pairwise differences between patients and controls were features characterizing H1–H2 variability (d = 0.74–0.88). To further illustrate these two discriminative features, Figure 3 shows simulated SPL and H1–H2 histograms representing the average patient and average control distributions. Simulated histograms were created using the "pearsrn" MATLAB function (MATLAB 2018, The MathWorks, Inc.) where random numbers were drawn from a distribution in the Pearson system with a mean, standard deviation, skewness, and kurtosis of the normal or patient data. Compared to an average control subject throughout a week, the average patient voiced with SPLs higher than their mean SPL for approximately 27 min longer and with H1–H2 values lower than their mean H1–H2 for approximately 20 min longer. Considering that all subjects averaged approximately 1 hr of phonation per day, 20–30 min of phonation represent nearly half of an entire day of voicing. Patients spending this large amount of time at higher vocal intensities with more abrupt vocal fold closure clearly reflect a phonatory behavior with a higher potential for phonotrauma. Also, a practical strength of using SPL skew is that it may be relatively immune to variability inherent in sensor placement and SPL calibration. For example, skew of the uncalibrated, neck-skin acceleration magnitude (in physical vibration units of dB cm/s^2) was correlated to SPL skew (r = .668) and still significantly different between patients and controls (d = 0.53, p < .001). Furthermore, a logistic regression model that substitutes the skewness of the uncalibrated ACC magnitude performs just as well as a model with SPL skew: total classification accuracy = 78.3% and AUC = 0.839 (95% CI [0.781, 0.898]).

Figure 3. Simulated data representing average patient (dashed lines) and matched healthy control (solid lines) weekly histograms of sound pressure level (SPL; black/left) and H1–H2 (gray/right).



As previously noted, SPL estimates are derived from a calibration procedure that determines the linear relationship between the amplitude of neck-skin acceleration and SPL 10–15 cm from the lips during a sustained vowel production (starting soft and ending loud). It is important to acknowledge that this relationship can vary by ± 5–6 dB during connected speech due to changes in the shape/occlusion of the supraglottal vocal tract (Švec et al., 2005) and can be randomly affected by measurement uncertainty (Bottalico et al., 2018). Thus, there is some uncertainty about the extent to which negative skewing of the ACC-based SPL distribution reflects comparable increases in the oral SPL (i.e., more frequent use of "louder" speech). However, irrespective of such uncertainties, the comparable results that were achieved in this study using the amplitude of the ACC signal calibrated to physical units of dB cm/s^2—and the correlation of subglottal pressure to ACC amplitude shown in other studies (Fryd et al., 2016)—support the view that patients with phonotrauma are employing higher laryngeal forces (including subglottal pressure) to phonate than healthy controls.

Since all of the patients in this study had vocal fold nodules or polyps during their week of ambulatory monitoring, it is not possible to empirically delineate which aspects of vocal behavior were present before the lesion formation (primary vocal hyperfunction) and which are in reaction to the presence of the lesions (secondary vocal hyperfunction; Verdolini et al., 2006). However, a negatively skewed SPL distribution could be hypothesized as a predisposing behavior for phonotraumatic lesion development. While

patients with vocal fold lesions commonly report difficulty talking softly (and a negative SPL distribution could be argued to result from avoiding soft talking), typical conversational levels rarely require this degree of reduced vocal intensity. Also, a negative SPL skew supports the clinical impression that patients with PVH talk louder than normal. However, as reflected in the present results, "louder than normal" may represent habitual tendencies to talk louder more often than average, instead of simply louder on average.

Since incomplete, hour-glass vocal fold closure patterns are commonly seen with phonotraumatic lesions during videostroboscopy (Colton et al., 1995), it may be surprising that patients did not vary as much toward higher values of H1–H2 (less abrupt and incomplete glottal closure) compared to their matched controls. A reasonable hypothesis could be that the patients were behaviorally compensating for their glottal gaps through hyperadduction, which would result in more abrupt and complete vocal fold closure. From a physiological perspective, reduced variance toward higher values of H1–H2 in the patient group could have partially resulted from their reduced variance toward higher f_o values. Higher f_o values are associated with more sinusoidal vocal fold kinematics and less overall glottal contact due to stretched lamina propria, which would also result in higher H1–H2. Although H1–H2 variability metrics and f_o variability metrics are correlated (mean Pearson r = .66), it seems unlikely that decreased f_o variability could solely account for the decreased H1–H2 variability as Cohen's d effect sizes are much larger for H1–H2 (d = 0.74–0.88) than f_o (d = 0.43–0.67).

It is possible that the diagnosis of vocal fold pathology and/or monitoring the patients could have affected their typical daily behavior, thus confounding the interpretation of results from any study using ambulatory monitoring (Hunter, 2012). A dramatic change in behavior due to these factors seems unlikely, especially since patients often need extensive voice therapy over the course of weeks or months to modify their habitual behaviors (Ziegler et al., 2014). Also, the majority of subjects reported forgetting that they were wearing the device. The data set contains a large number of professional and amateur vocalists, which may limit the study's external validity to patients with PVH who are not singers. Furthermore, it is possible that conflating singing and speech may have confounding influences on the results. Currently, we have developed an automatic singing detector to investigate the effect of singing on differences (and lack of differences) observed between patients and matched healthy controls (Ortiz et al., 2019).

We are currently monitoring these patients throughout treatment (both therapy and surgery) to enable comparisons of vocal function/behavior with and without lesions. Pre- versus postsurgery comparisons are especially important because, during postsurgical monitoring (prior to voice therapy), it is theoretically possible to observe the primary hyperfunctional behavior that caused the tissue damage without the potentially confounding influence of the lesions. Monitoring of behavioral changes that correlate with successful voice therapy after surgery has the potential to further verify which behaviors were most likely associated with the original causes of phonotrauma (primary hyperfunction).

Conclusion

Overall, compared to controls, the only highly discriminative differences in weekly voice use and vocal function of patients with phonotraumatic lesions (vocal fold nodules and polyps) are SPL skew and H1–H2 variability. In other words, patients tend to spend more time talking louder than average with more abrupt glottal closure compared to matched controls (and these two behaviors are not highly correlated among subjects). There seem to be small-to-medium significant differences in f_o variability, percent phonation time, and some phonatory/nonphonatory segments that are not primary contributors to classifying behavior associated with subjects who have phonotraumatic lesions and those who do not. More refined ambulatory measurements of hyperfunctional phonatory mechanisms, along with the examination of other potential contributing etiologic factors, are needed to improve the understanding of causative or associative risk factors for common phonotraumatic vocal fold lesions.

Acknowledgments

This work was supported by the Voice Health Institute and the National Institute on Deafness and Other Communication Disorders under Grants R33 DC011588 (Principal Investigator: Robert Hillman) and P50 DC015446 (Principal Investigator: Robert Hillman). The article's contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Institutes of Health.

The authors acknowledge the contributions of R. Petit for aid in designing and programming the smartphone application.

References

Astolfi, A., Carullo, A., Pavese, L., & Puglisi, G. E. (2015). Duration of voicing and silence periods of continuous speech in different acoustic environments. *The Journal of the Acoustical Society of America, 137*(2), 565–579.

Awan, S. N., Roy, N., Jetté, M. E., Meltzner, G. S., & Hillman, R. E. (2010). Quantifying dysphonia severity using a spectral/cepstral-based acoustic index: Comparisons with auditory–perceptual judgements from the CAPE-V. *Clinical Linguistics & Phonetics, 24*(9), 742–758.

Behlau, M., & Oliveira, G. (2009). Vocal hygiene for the voice professional. *Current Opinion in Otolaryngology & Head and Neck Surgery, 17*(3), 149–154.

Bottalico, P., & Astolfi, A. (2012). Investigations into vocal doses and parameters pertaining to primary school teachers in classrooms. *The Journal of the Acoustical Society of America, 131*(4), 2817–2827.

Bottalico, P., Astolfi, A., & Hunter, E. J. (2017). Teachers' voicing and silence periods during continuous speech in classrooms with different reverberation times. *The Journal of the Acoustical Society of America, 141*(1), EL26–EL31.

Bottalico, P., Graetzer, S., Astolfi, A., & Hunter, E. J. (2017). Silence and voicing accumulations in Italian primary school teachers with and without voice disorders. *Journal of Voice, 31*(2), 260.e11–260.e20.

Bottalico, P., Ipsaro Passione, I., Astolfi, A., Carullo, A., & Hunter, E. J. (2018). Accuracy of the quantities measured by four vocal dosimeters and its uncertainty. *The Journal of the Acoustical Society of America, 143*(3), 1591–1602.

Calosso, G., Puglisi, G. E., Astolfi, A., Castellana, A., Carullo, A., & Pellerey, F. (2017). A one-school year longitudinal study of secondary school teachers' voice parameters and the influence of classroom acoustics. *The Journal of the Acoustical Society of America, 142*(2), 1055–1066.

Carroll, T., Nix, J., Hunter, E., Emerich, K., Titze, I., & Abaza, M. (2006). Objective measurement of vocal fatigue in classical singers: A vocal dosimetry pilot study. *Otolaryngology—Head & Neck Surgery, 135*(4), 595–602.

Carullo, A., Vallan, A., & Astolfi, A. (2013). Design issues for a portable vocal analyzer. *IEEE Transactions on Instrumentation and Measurement, 62*(5), 1084–1093.

Carullo, A., Vallan, A., Astolfi, A., Pavese, L., & Puglisi, G. (2015). Validation of calibration procedures and uncertainty estimation of contact-microphone based vocal analyzers. *Measurement, 74,* 130–142.

Castellana, A., Carullo, A., Corbellini, S., & Astolfi, A. (2018). Discriminating pathological voice from healthy voice using cepstral peak prominence smoothed distribution in sustained vowel. *IEEE Transactions on Instrumentation and Measurement, 67*(3), 646–654.

Cheyne, H. A., Hanson, H. M., Genereux, R. P., Stevens, K. N., & Hillman, R. E. (2003). Development and testing of a portable vocal accumulator. *Journal of Speech, Language, and Hearing Research, 46*(6), 1457–1467.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.

Colton, R. H., Woo, P., Brewer, D. W., Griffin, B., & Casper, J. (1995). Stroboscopic signs associated with benign lesions of the vocal folds. *Journal of Voice, 9*(3), 312–325.

Cortés, J. P., Espinoza, V. M., Ghassemi, M., Mehta, D. D., Van Stan, J. H., Hillman, R. E., Guttag, J. V., & Zañartu, M. (2018). Ambulatory assessment of phonotraumatic vocal hyperfunction using glottal airflow measures estimated from neck-surface acceleration. *PLOS ONE, 13*(12), e0209017.

Espinoza, V. M., Zañartu, 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.

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.

Ghassemi, M., Van Stan, J. H., Mehta, D. D., Zañartu, M., Cheyne, H. A., II, Hillman, R. E., & Guttag, J. V. (2014). Learning to detect vocal hyperfunction from ambulatory neck-surface acceleration features: Initial results for vocal fold nodules. *IEEE Transactions on Biomedical Engineering, 61*(6), 1668–1675.

Goldman, S. L., Hargrave, J., Hillman, R. E., Holmberg, E., & Gress, C. (1996). Stress, anxiety, somatic complaints, and voice use in women with vocal nodules: Preliminary findings. *American Journal of Speech-Language Pathology, 5*(1), 44–54.

Henrich, N., d'Alessandro, C., & Doval, B. (2001). *Spectral correlates of voice open quotient and glottal flow asymmetry: Theory, limits and experimental data*. In P. Dalsgaard, B. Lindberg, H. Benner, & Z.-H. Tan (Eds.), EUROSPEECH 2001, Scandinavia, 7th European Conference on Speech Communication and Technology, 2nd INTERSPEECH Event, Aalborg, Denmark, September 3-7, 2001 (pp. 47–50). Aalborg, Denmark: Kommunik Grafiske Løsninger.

Herrington-Hall, B. L., Lee, L., Stemple, J. C., Niemi, K. R., & McHone, M. M. (1988). Description of laryngeal pathologies by age, sex, and occupation in a treatment-seeking sample. *Journal of Speech and Hearing Disorders, 53*(1), 57–64.

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.

Hillman, R. E., Heaton, J. T., Masaki, A., Zeitels, S. M., & Cheyne, H. A. (2006). Ambulatory monitoring of disordered voices. *Annals of Otology, Rhinology & Laryngology, 115*(11), 795–801.

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.

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.

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.

Holmberg, E. B., Hillman, R. E., Hammarberg, B., Sodersten, M., & Doyle, P. (2001). Efficacy of a behaviorally based voice therapy protocol for vocal nodules. *Journal of Voice, 15*(3), 395–412.

Hunter, E. J. (2012). Teacher response to ambulatory monitoring of voice. *Logopedics, Phoniatrics, Vocology, 37*(3), 133–135.

Hunter, E. J., & Titze, I. R. (2009). Quantifying vocal fatigue recovery: Dynamic vocal recovery trajectories after a vocal loading exercise. *Annals of Otology, Rhinology & Laryngology, 118*(6), 449–460.

Hunter, E. J., & Titze, I. R. (2010). Variations in intensity, fundamental frequency, and voicing for teachers in occupational versus nonoccupational settings. *Journal of Speech, Language, and Hearing Research, 53*(4), 862–875.

Karkos, P. D., & McCormick, M. (2009). The etiology of vocal fold nodules in adults. *Current Opinion in Otolaryngology & Head and Neck Surgery, 17*(6), 420–423.

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.

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

Kunduk, M., & McWhorter, A. J. (2009). True vocal fold nodules: The role of differential diagnosis. *Current Opinion in Otolaryngology & Head and Neck Surgery, 17*(6), 449–452.

Leonard, R. (2009). Voice therapy and vocal nodules in adults. *Current Opinion in Otolaryngology & Head and Neck Surgery, 17*(6), 453–457.

Lindstrom, F., Waye, K. P., Södersten, M., McAllister, A., & Ternström, S. (2011). Observations of the relationship between noise exposure and preschool teacher voice usage in day-care center environments. *Journal of Voice, 25*(2), 166–172.

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.

Maffei, M., Van Stan, J. H., Hillman, R. E., & Mehta, D. D. (2016). Correlating ambulatory voice measures with vocal fatigue self-ratings in individuals with MTD and normal controls. Paper presented at the Proceedings of Annual Convention of the American Speech-Language-Hearing Association. Philadelphia, PA, United States.

Masuda, T., Ikeda, Y., Manako, H., & Komiyama, S. (1993). Analysis of vocal abuse: Fluctuations in phonation time and intensity in 4 groups of speakers. *Acta Oto-Laryngologica, 113*(4), 547–552.

Mehta, D. D., Espinoza, V. M., Van Stan, J. H., Zañartu, M., & Hillman, R. E. (2019). The difference between first and second harmonic amplitudes correlates between glottal airflow and neck-surface accelerometer signals during phonation. *The Journal of the Acoustical Society of America, 145*(5), EL386–EL392.

Mehta, D. D., Van Stan, J. H., & Hillman, R. E. (2016). Relationships between vocal function measures derived from an acoustic microphone and a subglottal neck-surface accelerometer. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24*(4), 659–668.

Mehta, D. D., Van Stan, J. H., Zañartu, M., Ghassemi, M., Guttag, J. V., Espinoza, V. M., Cortés, J. P., Cheyne, H. A., II, & Hillman, R. E. (2015). Using ambulatory voice monitoring to investigate common voice disorders: Research update. *Frontiers in Bioengineering and Biotechnology, 3,* 155.

Mehta, D. D., Zañartu, M., Feng, S. W., Cheyne, H. A., II, & Hillman, R. E. (2012). Mobile voice health monitoring using a wearable accelerometer sensor and a smartphone platform. *IEEE Transactions on Biomedical Engineering, 59*(11), 3090–3096.

Morrow, S. L., & Connor, N. P. (2011). Comparison of voice-use profiles between elementary classroom and music teachers. *Journal of Voice, 25*(3), 367–372.

Nacci, A., Fattori, B., Mancini, V., Panicucci, E., Ursino, F., Cartaino, F., & Berrettini, S. (2013). The use and role of the Ambulatory Phonation Monitor (APM) in voice assessment. *Acta Otorhinolaryngologica Italica, 33*(1), 49–55.

Ortiz, A. J., Toles, L. E., Marks, K. L., Capobianco, S., Mehta, D. D., Hillman, R. E., & Van Stan, J. H. (2019). Automatic speech and singing classification in ambulatory recordings for normal and disordered voices. *The Journal of the Acoustical Society of America, 146*(1), EL22–EL27.

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

Popolo, P. S., Švec, J. G., & Titze, I. R. (2005). Adaptation of a pocket PC for use as a wearable voice dosimeter. *Journal of Speech, Language, and Hearing Research, 48*(4), 780–791.

Puglisi, G. E., Astolfi, A., Cantor Cutiva, L. C., & Carullo, A. (2017). Four-day-follow-up study on the voice monitoring of primary school teachers: Relationships with conversational task and classroom acoustics. *The Journal of the Acoustical Society of America, 141*(1), 441–452.

Roy, N., Gray, S. D., Simon, M., Dove, H., Corbin-Lewis, K., & Stemple, J. C. (2001). An evaluation of the effects of two treatment approaches for teachers with voice disorders: A prospective randomized clinical trial. *Journal of Speech, Language, and Hearing Research, 44*(2), 286–296.

Roy, N., Weinrich, B., Gray, S. D., Tanner, K., Toledo, S. W., Dove, H., Corbin-Lewis, K., & Stemple, J. C. (2002). Voice amplification versus vocal hygiene instruction for teachers with voice disorders: A treatment outcomes study. *Journal of Speech, Language, and Hearing Research, 45*(4), 625–638.

Searl, J., & Dietsch, A. (2014). Testing of the vocalog vocal monitor. *Journal of Voice, 28*(4), 523.e527–523.e537.

Södersten, M., Ternström, S., & Bohman, M. (2005). Loud speech in realistic environmental noise: Phonetogram data, perceptual voice quality, subjective ratings, and gender differences in healthy speakers. *Journal of Voice, 19*(1), 29–46.

Stevens, K. N. (1998). *Acoustic phonetics*. MIT Press.

Švec, J. G., Popolo, P. S., & Titze, I. R. (2003). Measurement of vocal doses in speech: Experimental procedure and signal processing. *Logopedics, Phoniatrics, Vocology, 28*(4), 181–192.

Švec, J. G., Titze, I. R., & Popolo, P. S. (2005). Estimation of sound pressure levels of voiced speech from skin vibration of the neck. *The Journal of the Acoustical Society of America, 117*(3), 1386–1394.

Swerts, M., & Veldhuis, R. (2001). The effect of speech melody on voice quality. *Speech Communication, 33*(4), 297–303.

Szabo, A., Hammarberg, B., Håkansson, A., & Södersten, M. (2001). A voice accumulator device: Evaluation based on studio and field recordings. *Logopedics, Phoniatrics, Vocology, 26*(3), 102–117.

Szabo Portela, A., Granqvist, S., Ternström, S., & Södersten, M. (2018). Vocal behavior in environmental noise: Comparisons between work and leisure conditions in women with work-related voice disorders and matched controls. *Journal of Voice, 32*(1), 126.e123–126.e138.

Titze, I. R., & Hunter, E. J. (2015). Comparison of vocal vibraion dose measures for potential damage risk criteria. *Journal of Speech, Language, and Hearing Research, 58*(5), 1425–1439.

Titze, I. R., Hunter, E. J., & Švec, J. G. (2007). Voicing and silence periods in daily and weekly vocalizations of teachers. *The Journal of the Acoustical Society of America, 121*(1), 469–478.

Titze, I. R., Švec, J. G., & Popolo, P. S. (2003). Vocal dose measures: Quantifying accumulated vibration exposure in vocal fold tissues. *Journal of Speech, Language, and Hearing Research, 46*(4), 919–932.

Van Stan, J. H., Gustafsson, J., Schalling, E., & Hillman, R. E. (2014). Direct comparison of three commercially available devices for voice ambulatory monitoring and biofeedback. *Perspectives on Voice and Voice Disorders, 24*(2), 80–86.

Van Stan, J. H., Maffei, M., Masson, M. L. V., Mehta, D. D., Burns, J. A., & Hillman, R. E. (2017). Self-ratings of vocal status in daily life: Reliability and validity for patients with vocal hyperfunction and a normative group. *American Journal of Speech-Language Pathology, 26*(4), 1167–1177.

Van Stan, J. H., Mehta, D. D., Zeitels, S. M., Burns, J. A., Barbu, A. M., & Hillman, R. E. (2015). Average ambulatory measures of sound pressure level, fundamental frequency, and vocal dose do not differ between adult females with phonotraumatic lesions and matched control subjects. *Annals of Otology, Rhinology & Laryngology, 124*(11), 864–874.

Verdolini, K., & Ramig, L. O. (2001). Review: Occupational risks for voice problems. *Logopedics, Phoniatrics, Vocology, 26*(1), 37–46.

Verdolini, K., Rosen, C., & Branski, R. C. (Eds.). (2006). *Classification manual for voice disorders-I, Special Interest Division 3,*

voice and Voice disorders, American Speech-Language Hearing Division. Erlbaum.

Zañartu, M., Galindo, G. E., Erath, B. D., Peterson, S. D., Wodicka, G. R., & Hillman, R. E. (2014). Modeling the effects of a posterior glottal opening on vocal fold dynamics with implications for vocal hyperfunctiona. *The Journal of the Acoustical Society of America, 136*(6), 3262–3271.

Zeitels, S. M., Hillman, R. E., Desloge, R., Mauri, M., & Doyle, P. (2002). Phonomicrosurgery in singers and performing artists: Treatment outcomes, management theories, and future directions. *Annals of Otology, Rhinology & Laryngology, 111*(12_suppl), 21–40.

Zhang, Z. (2016). Cause–effect relationship between vocal fold physiology and voice production in a three-dimensional phonation model. *The Journal of the Acoustical Society of America, 139*(4), 1493–1507.

Ziegler, A., Dastolfo, C., Hersan, R., Rosen, C. A., & Gartner-Schmidt, J. (2014). Perceptions of voice therapy from patients diagnosed with primary muscle tension dysphonia and benign mid-membranous vocal fold lesions. *Journal of Voice, 28*(6), 742–752.