ASHA American Speech-Language-Hearing Association

JSLHR

Research Article

Capturing the Intraspeaker Heterogeneity of Vocal Hyperfunction Using Spatiotemporal Indices of Relative Fundamental Frequency

Jenny Vojtech,[a] Laura E. Toles,[b] Daniel P. Buckley,[a,c] and Cara E. Stepp[a,c,d]

[a] Department of Speech, Language, & Hearing Sciences, Boston University, MA [b] Department of Otolaryngology—Head and Neck Surgery, The University of Texas Southwestern Medical Center, Dallas [c] Department of Otolaryngology—Head and Neck Surgery, Boston University School of Medicine, MA [d] Department of Biomedical Engineering, Boston University, MA

ARTICLE INFO

Article History:
Received February 25, 2025
Revision received April 16, 2025
Accepted May 23, 2025

Editor-in-Chief: Jessica E. Huber
Editor: Supraja Anand

https://doi.org/10.1044/2025_JSLHR-25-00138

ABSTRACT

Purpose: Hyperfunctional voice disorders are highly prevalent yet difficult to characterize objectively. Relative fundamental frequency (RFF) has the potential to characterize these disorders but faces limited clinical use due to intersubject variability in mean RFF values. This study examined whether RFF variability offers insights beyond traditional mean measures.

Method: Speech samples were collected from 132 adults: individuals with phonotraumatic vocal hyperfunction (PVH; *n* = 44), nonphonotraumatic vocal hyperfunction (NPVH; *n* = 44), and typical voices (controls; *n* = 44). Two measures of RFF variability—standard deviation and spatiotemporal index (STI)—were calculated along with mean RFF values. While standard deviation captures variability in magnitude, STI incorporates variability in time and magnitude. Permutational analyses of variance were conducted to assess relationships between group (PVH/NPVH/control) and the mean, standard deviation, and STI measures. Significant measures were entered along with demographic parameters into hierarchical multinomial logistic regression models using a training set (*n* = 102). Final model equations were then applied to an independent test set (*n* = 30) to predict group membership.

Results: Mean and STI measures showed significant group differences, whereas standard deviation did not. Both mean and STI measures improved model performance after adjusting for demographics. Receiver operating characteristic analysis on the test set yielded acceptable classification (area under curve = 0.78) for group membership.

Conclusions: Variability in RFF, especially when considering both time and magnitude, captures subtle features of vocal hyperfunction that may be overlooked by traditional mean measures. These findings underscore the clinical value of advanced RFF variability metrics in characterizing vocal hyperfunction.

Supplemental Material: https://doi.org/10.23641/asha.29903054

Vocal hyperfunction (VH) is a prevalent voice condition that presents in approximately half of all cases referred to multidisciplinary voice clinics (Dworkin-Valenti et al., 2018; Roy, 2003). Clinical efforts to characterize VH point to abnormally increased muscle activity in the larynx as the primary factor underlying clinical symptoms such as dysphonic voice quality and, more commonly, persistent complaints of vocal effort that arise in cases of VH (Colton et al., 2006; Hanschmann et al., 2010; Jiang & Titze, 1994; van Mersbergen et al., 2021). Hyperfunctional behaviors can be triggered by different etiological factors—such as patterns of voice misuse (Altman et al., 2005; Van Houtte et al., 2010) or altered laryngeal biomechanics (Espinoza et al., 2017)—and can manifest on their own or as a compensatory response to vocal injury (Hillman et al., 2020). In isolation, VH is referred to as "non-phonotraumatic" (nonphonotraumatic

Correspondence to Jenny Vojtech: jmvo@bu.edu. ***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

vocal hyperfunction [NPVH]) and is commonly diagnosed as primary muscle tension dysphonia. When vocal fold lesions such as polyps or nodules are present, compensatory hyperfunctional behaviors are classified as "phonotraumatic" (phonotraumatic vocal hyperfunction [PVH]), or secondary muscle tension dysphonia (Hillman et al., 2020).

Although VH is recognized as a prominent feature of many voice disorders (Bhattacharyya, 2014), clinical assessment of the underlying muscle activity remains a challenge. Current methods to evaluate laryngeal muscle activity typically include patient-reported outcomes—such as the Vocal Fatigue Index (Nanjundeswaran et al., 2015), Voice Handicap Index (Jacobson et al., 1997), Vocal Tract Discomfort Scale (Mathieson et al., 2009), or Voice-Related Quality of Life questionnaire (Hogikyan & Sethuraman, 1999)—auditory-perceptual judgments of vocal strain, manual palpation of the perilaryngeal region, and visual-perceptual judgments via laryngoscopy. These tools provide substantial insight into overall vocal function and health, but their subjective nature can introduce rater error and bias, and the clinical heterogeneity of VH often makes them unreliable (Askenfelt & Hammarberg, 1986; Dejonckere et al., 1996; Jafari et al., 2017; Lowell et al., 2012; Thomas et al., 2023; Zraick et al., 2011). As a result, many researchers have turned to developing objective acoustic measures to supplement existing assessment methods. For instance, cepstral peak prominence has been demonstrated to be an effective indicator of dysphonia severity (Maryn et al., 2009) and is therefore considered as an objective complement to auditory perceptions of voice quality (Patel et al., 2018).

Clinical Heterogeneity of VH

Although objective acoustic measures may help mitigate some aspects of subjective error and bias inherent in auditory-perceptual evaluations—such as variability in clinician experience, listener fatigue, or differing interpretations of perceptual scales—their utility in assessing VH is still limited. This is likely due, in part, to the extreme heterogeneity of the condition: It can present with (PVH) or without (NPVH) organic lesions, and the overall severity of dysphonia can vary widely from normal to severe (McDowell et al., 2022; Morrison, 1997; Roy, 2008). Additionally, VH is often characterized by patient complaints of vocal effort during speech (Aronson & Bless, 2009; Dworkin et al., 2000; Koufman & Blalock, 1991; Morrison et al., 1983; Roy, 2008). Acoustic measures objectively capture features of the voice signal, such as periodicity, spectral tilt, and harmonic structure; however, these reflect the *output* of phonation and may not directly map onto the underlying *input* mechanisms of VH (i.e., dysregulated laryngeal tension). Furthermore, many perceptual signs associated with VH-like strain are not pathognomonic and may be present in a range of voice disorders. As such, no single acoustic measure consistently demonstrated sufficient sensitivity and specificity to reliably detect or monitor PVH or NPVH, complicating efforts to establish objective baselines or track progress during behavioral voice therapy.

Recent investigations indicate that VH is a product of disrupted sensorimotor control, where deficiencies in using auditory feedback to update predefined motor commands over time lead to a spectrum of voice changes (e.g., dysphonia) and symptoms (e.g., vocal effort) observed during clinical assessments. Studies using altered auditory feedback paradigms have indicated that these deficiencies often result in larger individual variability in vocal responses, reflecting inaccuracies in feedforward control (Abur et al., 2021; Stepp et al., 2017). Evidence of increased within-speaker variability in VH is further reinforced by studies examining multiple speech subsystems using voice acoustics (Belsky et al., 2021; McKenna et al., 2020), aerodynamics (Belsky et al., 2021; Gillespie et al., 2013; Higgins et al., 1999), and articulatory kinematics (Shipurkar et al., 2023). In cases of PVH, vocal fold lesions appear to contribute to an inherently unstable physical system (Crocker et al., 2024; Free et al., 2023), which may increase the reliance on feedback control mechanisms. The combination of feedforward control inaccuracies with this additional, increased demand for feedback regulation is likely to result in even greater variability and less efficient movement patterns during speech production. Consequently, individuals with PVH may exhibit greater variability in their speech production compared to those with NPVH.

The heterogeneous clinical presentation of VH, along with consistent findings of variability across multiple speech subsystems, suggests that both between- and within-speaker variability may be characteristics of the condition rather than obstacles to accurate assessment. Viewing variability as a defining feature rather than a limitation could pave the way for more sensitive and specific tools that leverage variability measures to better capture the spectrum of VH presentation.

Relative Fundamental Frequency for Characterizing VH

Over the past decade, relative fundamental frequency (RFF) has emerged as a promising acoustic measure for characterizing VH. RFF is calculated as the relative change in voice fundamental frequency ($F0$) as an individual begins or ends voicing. The rationale for using RFF stems from the foundational relationships among RFF, voice $F0$, and the biomechanical properties of the vocal folds: RFF captures instantaneous changes in voice

*F*0 during voice offsets or onsets, which requires adjustments in the vocal fold properties (i.e., length, mass, and tension) to properly regulate the termination or initiation of vocal fold vibration (i.e., phonation; van den Berg, 1958). The RFF values derived from instantaneous *F*0 closest to the points of voice offset and onset—known as "RFF_{offset10}" and "RFF_{onset1}," respectively[1]—are the most responsive to changes in laryngeal muscle activity (Goberman & Blomgren, 2008). This responsiveness is evidenced by clinically significant changes in the magnitudes of RFF_{offset10} and RFF_{onset1} over the course of behavioral voice therapy for individuals with VH (Stepp et al., 2011), as well as statistically significant differences between individuals with and without VH (Stepp et al., 2010).

Research suggests that adults without VH exhibit the highest mean values for RFF_{offset10} and RFF_{onset1}, followed by adults with NPVH and then those with PVH (e.g., Ferrán et al., 2024; Heller Murray et al., 2017; Roy et al., 2016; Stepp et al., 2010, 2012). This hierarchy is thought to reflect differences in baseline laryngeal tension among these groups. The broad schema is that *F*0 increases with increased laryngeal tension, whether direct (i.e., increased contraction of the thyroarytenoid) or indirect (e.g., cricothyroid activation, laryngeal elevation). Evidence suggests that laryngeal tension transiently elevates to increase vocal fold stiffness during voice offsets (Stevens, 1977; Watson, 1998) and onsets (Löfqvist et al., 1989; Stevens, 1977). In individuals with VH, particularly those with PVH, it has been hypothesized that increased baseline laryngeal tension may limit the use of additional tension as a strategy for devoicing (voice offset) and reinitiating voicing (voice onset).

Despite efforts to validate RFF as a potential tool for clinical voice assessments, considerable variability in mean RFF across individuals has hindered its widespread adoption in clinical settings. Heller Murray et al. (2017) investigated the ability of RFF_{offset10} and RFF_{onset1} to distinguish between VH subtypes (PVH and NPVH) from individuals without VH (controls) using receiver operating characteristic (ROC) curves. The authors identified a cutoff threshold for classifying VH subtypes by maximizing the positive likelihood ratio to reflect the best balance between detecting positive cases (PVH or NPVH) while minimizing false positives. They observed low sensitivity and high specificity in distinguishing controls from either PVH (sensitivity = 0.43, specificity = 0.98) or NPVH (sensitivity = 0.19, specificity = 0.91). The high specificity observed for discriminating VH indicates that the model was very good at correctly identifying individuals who did not have VH, but at the cost of low sensitivity. This low sensitivity indicates that many individuals who actually had PVH or NPVH were missed, resulting in a significant number of false negatives. One common metric used to evaluate diagnostic accuracy is the Youden Index (J = sensitivity + specificity − 1), which reflects a method's ability to balance sensitivity and specificity; values above 0.5 (no discriminatory power) are generally considered to indicate acceptable performance. The results from Heller Murray et al. (2017) translate to relatively low diagnostic accuracy overall; for instance, the Youden Index for these methods was 0.41 for PVH and just 0.10 for NPVH, both of which fall below the 0.5 threshold often considered acceptable. This suggests that—while the approach was effective at reducing false positives—it was not as effective at accurately identifying those with VH, which limits its practical application in clinical settings where detecting all cases is crucial. These findings highlight the limitations of relying solely on mean RFF values for classification.

More recent efforts from Kapsner-Smith et al. (2022) demonstrated substantial improvements toward using RFF for assessing VH. They demonstrated that incorporating more sophisticated RFF metrics (such as the difference between RFF_{onset1} and RFF_{onset6}) could be combined with measures of smoothed cepstral peak prominence (CPPS) to improve discriminatory power. The authors incorporated two thresholding methods to assess the potential of their model as a screening tool for VH. The first method used a fixed cutoff of 0.50, resulting in a relatively higher sensitivity for discriminating PVH (sensitivity = 0.72, specificity = 0.78) or NPVH (sensitivity = 0.66, specificity = 0.66) from controls. Their second method aimed to maximize sensitivity while maintaining a specificity of at least 0.60, following the approach outlined by Awan et al. (2016); here, they observed even higher sensitivities for distinguishing PVH (sensitivity = 0.90, specificity = 0.60) or NPVH (sensitivity = 0.68, specificity = 0.60) from controls. Still, the diagnostic accuracy of these methods—reflected by Youden Indices of 0.50 for PVH in both studies—remains modest, indicating just acceptable performance. Clarifying the limited sensitivity–specificity trade-offs of these methods is important for contextualizing their clinical utility and highlights the continued need for more robust VH classification tools.

By optimizing the cutoff to maximize sensitivity, the authors were able to significantly improve the detection of VH subtypes, even at the cost of reduced specificity. This approach resulted in higher sensitivities for PVH, suggesting that a focus on sensitivity—ensuring that more true

[1] RFF is the semitone difference between the *F*0 of a specific voicing cycle and a (pseudo)steady-state *F*0 within a vowel–voiceless consonant–vowel (VCV) sequence. The last 10 cycles before the voiceless consonant (offset) and the first 10 cycles after it (onset) are used. The (pseudo) steady-state reference cycle is closest to the midpoint of each vowel (i.e., Offset Cycle 1 and Onset Cycle 10). For instance, RFF_{offset10} is the semitone difference between the 10th offset cycle and the reference cycle (Offset cycle 1).

cases are identified—can be crucial for screening tools, even if it means accepting a higher rate of false positives. The combination of RFF metrics with CPPS measures further enhanced sensitivity, demonstrating that adjusting thresholds and coupling CPPS with complex RFF measures can improve the detection of VH subtypes more effectively than relying solely on a fixed cutoff or using mean RFF_{offset10} and RFF_{onset1} values alone.

Nevertheless, applying this approach in clinical practice may present some challenges. Cepstral measures are highly sensitive to the recording environment and microphone specifications (Patel et al., 2018), requiring consideration of factors such as ambient noise levels, microphone type and placement, and signal processing methods to ensure accurate and reliable CPPS measurements. Moreover, separate software is required to compute RFF (MATLAB) and CPPS (Praat or ADSV), which adds complexity and necessitates further training and resources for clinical use. Taken together, these findings emphasize the need for a balanced approach that prioritizes both diagnostic accuracy and practical feasibility in clinical settings.

Perspectives on Variability in RFF

As prior research suggests that disrupted sensorimotor control in VH may be linked to increased variability in speech production, it is not unexpected that mean RFF values offer limited discrimination ability on their own. However, this variability has traditionally been viewed as a limitation, particularly regarding the clinical applicability of RFF (Roy et al., 2016). Few studies have specifically examined the within-speaker variability of RFF, and most have focused on minimizing its influence. For example, Groll et al. (2022) observed that this variability could arise from limitations in current RFF algorithms, which rely on microphone acoustics. These acoustics may not accurately capture the physiological points of phonatory initiation or termination, such as the end of vocal fold contact at voice offset. Yet, the presence of variability itself may provide valuable insight into the underlying nature of VH. Rather than treating this variability as an obstacle to accurate measurement, we propose that it might be useful for improving the discriminative power of RFF in VH assessment.

To explore this concept, we investigated two distinct measures of within-speaker variability: a basic measure (standard deviation) and a more complex measure known as the spatiotemporal index (STI). As RFF is a measure of how instantaneous *F*0 changes as someone stops or starts voicing, it is parallel to STI, which assesses movement pattern stability exhibited over repeated performance of the same motor task (such as speech production; Smith et al., 2000). By applying standard deviation or STI to examine an individual's RFF estimates across their VCV productions, we aimed to determine whether incorporating measures of variability can enhance our ability to discriminate between VH subtypes and individuals with typical voices, potentially capturing aspects of the condition that mean RFF values alone may miss.

Thus, the purpose of this study was to compare the effectiveness of a simple measure of variability (standard deviation) with a more complex measure (STI) in improving group discrimination and to explore whether incorporating measures of RFF variability can enhance the diagnostic utility of RFF in identifying and differentiating VH subtypes. As such, we hypothesized that measures of variability capture important aspects of VH not reflected by mean RFF values alone, where entering variability measures into a classification model that includes mean RFF values (RFF_{offset10}, RFF_{onset1}) will significantly enhance the model's ability to discriminate groups.

As a secondary, exploratory aim, we also sought to examine how RFF variability may differ between VH subtypes and controls. We suspect that while mean RFF values reflect overall levels of laryngeal tension during phonation, variability in RFF (as captured by standard deviation or STI) may provide insight into the stability or efficiency of motor planning processes. Although we cannot directly measure laryngeal physiology or motor control using acoustic estimates of RFF, we sought to indirectly assess group-level distinctions in sensorimotor function through these acoustic proxies. We hypothesized that VH is characterized by greater variability in laryngeal muscle activity during voice offsets and onsets. Specifically, we expected that individuals with VH (PVH, NPVH) would exhibit significantly greater variability in RFF values (via higher standard deviation and STI values) than age- and sex-matched controls and, moreover, that individuals with PVH would exhibit significantly greater variability in RFF values than individuals with NPVH.

Method

Participant Characteristics

Speech recordings were collected from adults, comprising three groups: individuals with PVH, individuals with NPVH, and controls with typical voices. All participants completed written consent in compliance with the Boston University Institutional Review Board (Protocol #2625), University of Washington Institutional Review Board (Protocol #36181), or The University of Texas Southwestern Institutional Review Board (Protocol #STU-2022-0388).

The control group had no self-reported history of speech, language, or hearing disorders and was age- and sex-matched to both VH groups within a 5-year range. Eligibility criteria for the VH groups required a diagnosis consistent with PVH—defined as structural pathology (e.g., nodules, polyp) accompanied by VH, as determined by a referring otolaryngologist—or NPVH, defined as primary muscle tension dysphonia and confirmed by a referring laryngologist. For the NPVH group, inclusion criteria were such that individuals presented with supraglottic compression and lack of observable structural or neurological pathology. Exclusion criteria for the study encompassed neurogenic diagnoses (e.g., laryngeal dystonia, vocal fold paresis/paralysis) and structural pathologies inconsistent with PVH (e.g., papilloma, carcinoma). These criteria resulted in a data set of 44 controls (35 women, nine men; M_{age} = 38.8 years, SD_{age} = 16.6 years, range: 18–77 years), 44 individuals with PVH (35 women, nine men; M_{age} = 38.9 years, SD_{age} = 16.5 years, range: 18–73 years), and 44 individuals with NPVH (35 women, nine men; M_{age} = 39.5 years, SD_{age} = 16.6 years, range: 18–74 years), comprising a total of 132 participants.

The auditory-perceptual quality of each voice was evaluated by a certified speech-language pathologist specializing in voice disorders. The speech-language pathologist, who was blinded to the study conditions, completed the 100-mm visual analog scale for overall severity of dysphonia from the Consensus Auditory-Perceptual Evaluation of Voice using the nonlinearly placed textual severity labels as originally published (American Speech-Language-Hearing Association, 2002). Mean overall severity of dysphonia scores were as follows: 7.3 (*SD* = 7.7, range: 0–30.3) for the control group, 20.8 (*SD* = 13.2, range: 0–51.0) for the NPVH group, and 29.1 (*SD* = 17.2, range: 6.0–66.0) for the PVH group.

Acoustic Recording Procedures

All signals were acquired digitally and analysis occurred offline. Participants were recorded in a waiting area or quiet room at one of four locations: (a) Boston Medical Center with a dynamic headset microphone (*n* = 14; Shure WH20XLR), (b) Boston University with a condenser headset microphone (*n* = 15; Shure SM35XLR), (c) University of Washington with a dynamic headset microphone (*n* = 8; Shure WH20XLR), or (d) UT Southwestern Medical Center with a condenser headset microphone (*n* = 95; AKG C520).[2]

Acoustic signals were sampled at 44.1 kHz with 16-bit resolution. As in prior RFF studies, participants produced three different VCV tokens with the voiceless consonant /f/ and the vowels /ɑ/, /i/, and /u/ (Lien et al., 2014). VCVs were modeled by the examiner with equal stress on each syllable before being repeated by participants at their typical pitch and loudness between 3 and 8 times (e.g., /ɑfɑ ɑfɑ ɑfɑ ifi fi ifi ufu ufu ufu/). The number of VCV repetitions varied across sites due to protocol differences at each recording location.

Data Processing

Recordings from each participant were processed in MATLAB 23.2 (MathWorks) using an open-source, semi-automated algorithm called *aRFF-AP*[3] to generate RFF traces for each VCV. Using *aRFF*-AP, the first author (J.V.) visually confirmed the location of the voiceless consonant in each VCV. The algorithm then calculated RFF and automatically rejected any RFF traces that failed to meet a set of predetermined criteria such as glottalization, misarticulation, or voicing the consonant. Resulting RFF traces comprised 20 values: 10 values for voice offset and 10 values for voice onset. At least two valid RFF traces were available for all participants (valid voice offset traces per participant: *M* = 9.0, *SD* = 5.0, range: 2–23; valid voice onset traces per participant: *M* = 9.8, *SD* = 5.0, range: 2–23).[4]

The Voice Offset Cycle 10 and Voice Onset Cycle 1 values of each available RFF trace for each speaker were averaged to produce mean RFF_{offset10} and RFF_{onset1} values, termed M_{offset10} and onset M_{onset1}, respectively. To compute within-speaker variability in RFF, two methods were carried out using custom scripts in MATLAB (Version 23.2), as described below.

Standard Deviation of RFF

For each participant, we calculated the standard deviation of available RFF traces at voice offset (SD_{offset10}) and onset (SD_{onset1}). Because most participants had fewer than 30 RFF traces, we applied Bessel's correction (using $n - 1$ instead of n in the formula) to account for the bias that can occur when estimating population parameters from small samples and, in turn, ensure a more accurate estimate of the standard deviation.

STI of RFF

For each participant, we also calculated the STI across RFF traces at voice offset (STI_{offset}) and onset (STI_{onset}) as follows:

[2] Relationships between microphone type and outcome measures were not statistically significant (all $p > .05$), as detailed in Supplemental Material S1, Table S1. Regardless, we acknowledge differences in microphone characteristics may still represent a potential source of uncontrolled variability in the study.

[3] Available from https://sites.bu.edu/stepplab/research/rff/.

[4] Relationships between the number of RFF traces and the outcome measures were not statistically significant (all $p > .05$), as detailed in Supplemental Material S1, Table S2.

1. Standardization: Each 10-point RFF trace was normalized by converting the RFF values into *z* scores, thereby centering them around the mean and scaling to the standard deviation of the trace.

2. Resample and extract values: Each standardized RFF trace was resampled to 1,000 points using a cubic spline procedure to interpolate between data points. Then, values were taken at 2% intervals to yield a trace with 50 evenly spaced points, following the method from Smith et al. (2000). This final standardized and time-normalized trace allowed us to make direct comparisons across participants.

From here, the sample standard deviation was computed across RFF traces at each of these 50 points. The resulting standard deviation vector was summed, and a gamma-based correction was applied to address potential bias due to the number of traces, as in Wisler et al. (2022). This resulted in a one STI estimate at voice offset and one at voice onset per speaker. Figure 1 shows a schematic overview of the process for computing STI. The result of all data processing was six measures for each participant: M_{offset10}, M_{onset1}, SD_{offset10}, SD_{onset1}, STI_{offset}, and STI_{onset}.

Statistical Analysis

All statistical analyses were conducted using Python (Version 3.8) software.

Multivariate Analyses

A permutational multivariate analysis of variance (PERMANOVA) was conducted to examine how the three groups (control, PVH, NPVH) differed on each of the six RFF measures (M_{offset10}, M_{onset1}, SD_{offset10}, SD_{onset1}, STI_{offset}, and STI_{onset}) using the *scikit-bio* (Version 0.6.2) and *SciPy* (Version 1.9.3) packages. The nonparametric PERMANOVA was selected over a traditional parametric multivariate analysis of variance due to violations of the assumption of multivariate normality (Anderson, 2001, 2017; McArdle & Anderson, 2001), despite the data conforming to Box's *M* test for homogeneity of covariance matrices. Significance was set a priori at $p < .05$ and effect sizes were quantified using partial eta-squared curvilinear correlations (η_p^2) with interpretations based on criteria from Cohen (1988). Euclidean distance was used to compute the distance matrix, producing an *F* statistic that is directly comparable to the traditional parametric *F* statistic (Anderson, 2001).

Significant multivariate effects were further analyzed using permutational analyses of variance (ANOVAs) with Bonferroni correction for each RFF measure to identify contributing variables. As a result of the Bonferroni correction, resulting *p* values were adjusted to account for multiple comparisons, which decreased the significance thresholds to reduce the risk of Type I errors. Post hoc permutational pairwise comparisons with Bonferroni correction were subsequently conducted on significant RFF measures. Cohen's *d* was used to measure effect sizes for pairwise comparisons. All permutational analyses were performed with 10,000 permutations to balance statistical precision and computational efficiency.

Regression Model Construction

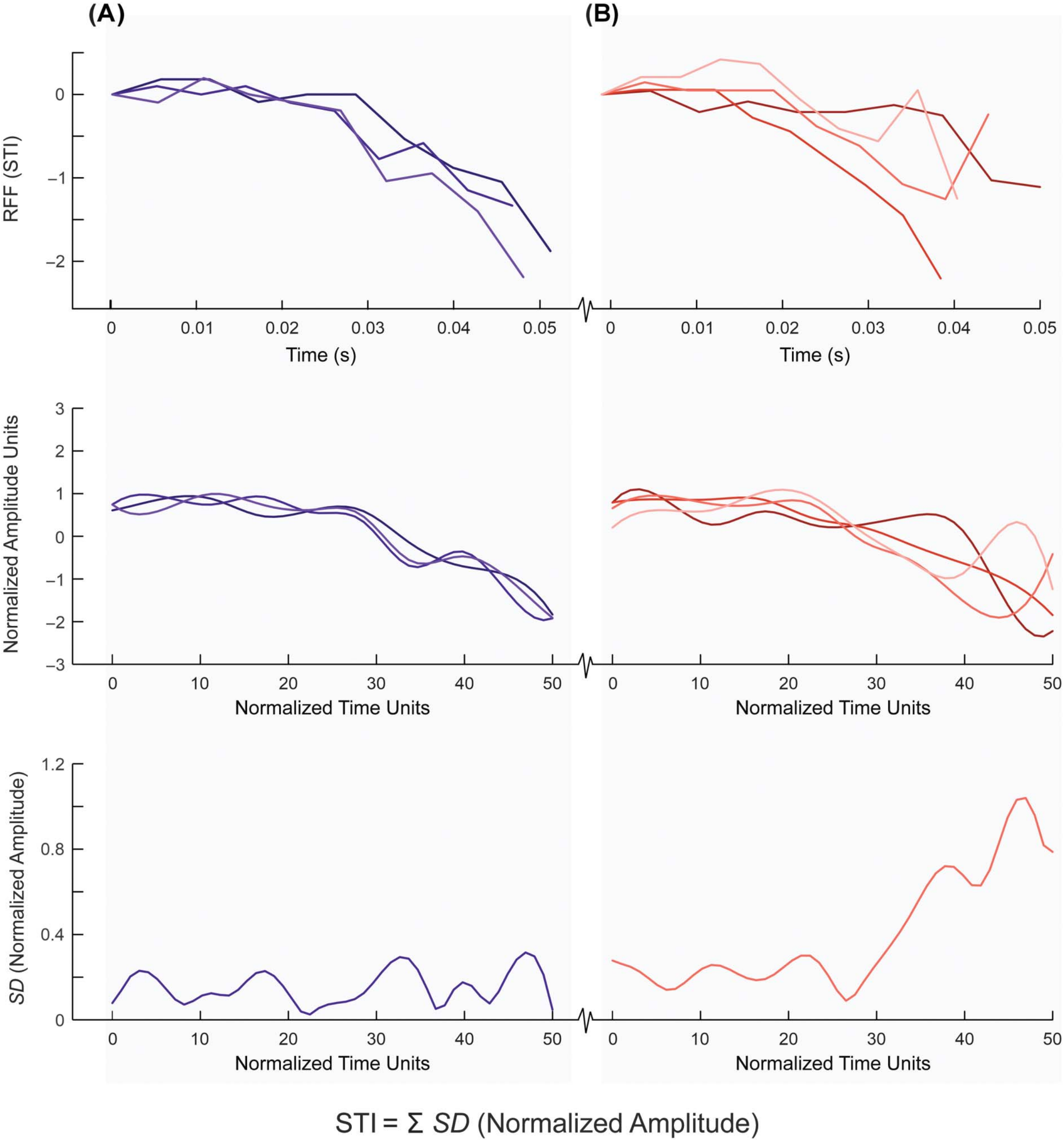
Data splitting and normalization. Hierarchical multinomial logistic regressions were performed using only the RFF variability measures that showed statistically significant differences in the PERMANOVA to evaluate their impact on predicting group membership. The 132-participant data set was randomly split into training (78%) and test (22%) sets using sex and group as stratification factors. This split yielded 102 participants in the training set (n = 34 per group) and 30 participants in the independent test set (n = 10 per group). RFF measure data were normalized across the training data set using a quantile transformation via the *QuantileTransformer* function of the *scikit-learn* package (Version 1.3.2; Pedregosa et al., 2011) to approximate a normal distribution. This transformation was subsequently applied to the test data set for downstream model performance evaluations.

Model design. Regression models were constructed for each statistically significant RFF variability measure (standard deviation or STI) using the MNLogit function from the *statsmodel* package (Version 0.13.2). Each model included age and sex as baseline predictors, followed by mean RFF measures (*M*), and then an RFF variability measure. Resulting model fits were characterized using Nagelkerke pseudo-*R*-squared (pseudo-R^2; Nagelkerke, 1991) estimates and compared using the change in chi-square (χ^2_{change}).

Model Evaluation and Performance

Training data set evaluation. One-vs-rest ROC curve analyses were performed on the training data set using *scikit-learn* to determine cutoff thresholds that balanced sensitivity and specificity by minimizing the Euclidean distance to the ideal point (0, 1) in ROC space. To enhance reliability, bootstrap resampling was used to generate 10,000 samples by randomly sampling with replacement from the original training data set, maintaining the same sample size and class proportions in each replicate. Model fits on the training data set—including regression coefficients, pseudo-R^2, χ^2_{change}, and ROC metrics (area under the curve [AUC], sensitivity, and specificity)—were characterized using median and 95% Wilson confidence intervals across bootstrap samples.

Figure 1. Example of two participants (A and B), showing how the spatiotemporal index (STI) is calculated at voice offset. Relative fundamental frequency (RFF) traces (top row) first undergo time and amplitude normalization (middle row) and are then aggregated across normalized time by computing the standard deviation at each time unit (bottom row). STI is derived by summing these standard deviation values.



Test data set performance. For models involving RFF variability, the regression equations derived from each bootstrap sample were applied to the independent test set to predict group membership. Class probabilities were aggregated using medians, and then one-vs-rest ROC analyses were performed to evaluate discriminative performance on the test data set. Median Euclidean thresholds from the bootstrap samples (constructed from the training data) were applied to test set probabilities to produce final performance estimates. Final metrics of AUC, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were calculated for each group to evaluate model performance. While AUC, sensitivity, and specificity evaluate test accuracy across all possible outcomes, the inclusion of PPV and NPV offers additional insights into the predictive value of test results based on the condition's prevalence in the population. PPV measures the proportion of true positive results among all positive predictions, indicating the likelihood that a positive test result correctly identifies a condition. Conversely, NPV measures the proportion of true negative results among all negative predictions, reflecting the likelihood that a negative test result correctly excludes the condition. By using these five metrics, we aim to provide a more detailed description of final model performance in predicting outcomes.

Results

Multivariate Analyses

Group data for each measure are presented in Figure 2 for descriptive purposes. The PERMANOVA revealed significant differences between groups with a pseudo-$F(2, 129) = 5.215$ and associated $p < .001$. Follow-up permutational ANOVA results are shown in Table 1. Group showed a significant large effect on M_{offset10}, $F(2, 129) = 10.147$, $p = .001$, and a significant, medium effect on M_{onset1}, $F(2, 129) = 7.485$, $p = .003$, but was not significant in the models for SD_{offset10} or SD_{onset1}. Group was significant for both STI measures, revealing significant medium effects for both for STI_{offset}, $F(2, 129) = 5.063$, $p = .005$, and STI_{onset}, $F(2, 129) = 5.364$, $p = .008$.

Post hoc pairwise comparisons revealed several significant differences in mean and STI measures across groups. For M_{offset10}, the PVH group exhibited significantly lower values than the control ($p = .006$, $d = 0.71$) and NPVH ($p < .001$, $d = 0.92$) groups. For M_{onset1}, both NPVH ($p = .010$, $d = 0.64$) and PVH ($p = .002$, $d = 0.79$) groups demonstrated significantly lower values than the control group. For spatiotemporal indices, STI_{offset} values were significantly higher in the NPVH group compared to the PVH group ($p = .007$, $d = 0.66$), and STI_{onset} values were significantly lower in the control group compared to PVH ($p = .020$, $d = -0.62$) and NPVH ($p = .014$, $d = -0.63$) groups.

Regression Model Construction

Based on these results, a single hierarchical multinomial logistic regression model was constructed, incorporating the mean and STI measures that showed significant group effects. One regression model was fit to the training data set to determine the impact of introducing STI at voice offset (STI_{offset}) and onset (STI_{onset}) as predictor variables for discriminating control, PVH, and NPVH groups. The order of entry of predictor parameters included (a) demographic parameters, (b) mean parameters, and (c) STI parameters. The result of this model is shown in Table 2.

Figure 2. Group averages for (a) mean, (b) standard deviation, and (c) spatiotemporal index (STI) of relative fundamental frequency (RFF) for voice offset (left) and onset (right). Error bars show 95% confidence intervals. *p < .05. NPVH = nonphonotraumatic vocal hyperfunction; PVH = phonotraumatic vocal hyperfunction; nu = no units.

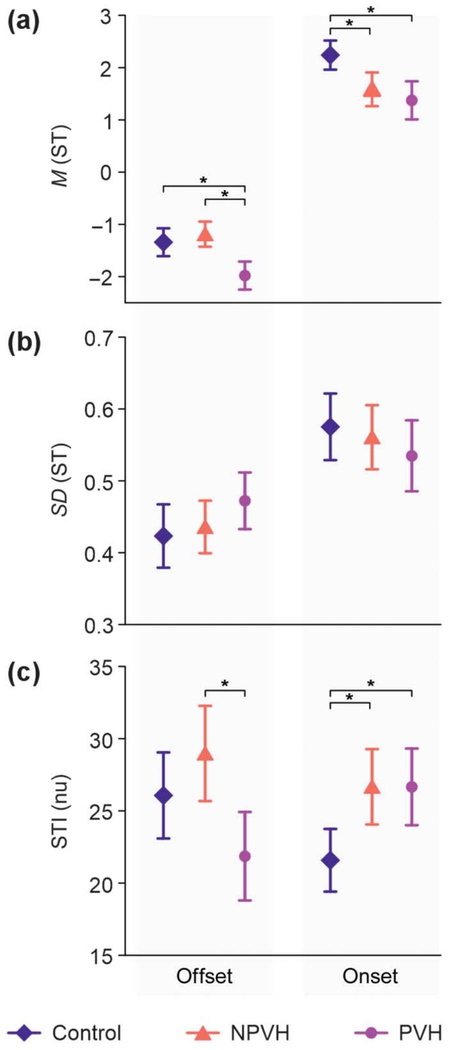


Table 1. Permutational analysis of variance for relative fundamental frequency (RFF) measures by group.

Measure	F	df	p	η_p^2	Effect size interpretation
M_{offset10}	10.147	2, 129	**.001**	.14	Large
SD_{offset10}	1.545	2, 129	.238	.02	Small
STI_{offset}	5.063	2, 129	**.005**	.07	Medium
M_{onset1}	7.485	2, 129	**.003**	.10	Medium
SD_{onset1}	0.732	2, 129	.488	.01	Small
STI_{onset}	5.364	2, 129	**.008**	.08	Medium

Note. Bold values indicate statistically significant measures. STI = spatiotemporal index.

Table 2. Hierarchical multinomial logistic regression model on the training data set.

Model	Predictors	χ^2_{change}	*p*	*Pseudo-R^2_{adj}*
Block 1	Age, sex	3.62 (0.52, 11.95)	.164 (.003, .770)	.04 (.00, .28)
Block 2	**Age, sex, M_{offset10}, M_{onset1}**	**27.29 (11.85, 48.75)**	**< .001 (.000, .003)**	**.59 (.30, .80)**
Block 3	**Age, sex, M_{offset10}, M_{onset1}, STI_{offset}, STI_{onset}**	**9.00 (1.98, 21.40)**	**.011 (.000, .372)**	**.68 (.41, .86)**

Note. All values are provided as median (95% confidence interval) across 1,000 bootstrap samples, and bolded values indicate statistical significance (*p* < .05). STI = spatiotemporal index.

$$\log\left(\frac{p(\text{NPVH})}{p(\text{Control})}\right) = 0.246 - 0.082 \times \text{Age} + 0.042 \times \text{Sex(Male)} - 0.379 \times M_{\text{offset10}} - 0.810 \times M_{\text{onset1}} + 0.709 \times STI_{\text{offset}} + 0.009 \times STI_{\text{onset}} \quad (1)$$

$$\log\left(\frac{p(\text{PVH})}{p(\text{Control})}\right) = 0.186 - 0.123 \times \text{Age} + 0.062 \times \text{Sex(Male)} - 1.374 \times M_{\text{offset10}} - 0.853 \times M_{\text{onset1}} + 0.814 \times STI_{\text{offset}} - 0.002 \times STI_{\text{onset}} \quad (2)$$

Table 3. Logistic regression results on the training data set, shown as medians with (95% Wilson confidence interval) from 10,000 bootstrap samples.

Model	*b*
Block 1	
NPVH–control	
Intercept	0.000 (–0.663, 0.680)
Age	–0.035 (–0.475, 0.493)
Sex (male)	0.000 (–0.136, 0.129)
PVH–control	
Intercept	–0.003 (–0.677, 0.653)
Age	0.031 (–0.423, 0.526)
Sex (male)	–0.004 (–0.139, 0.128)
Block 2	
NPVH–control	
Intercept	0.240 (–0.513, 1.117)
Age	–0.003 (–0.506, 0.527)
Sex (male)	0.053 (–0.096, 0.220)
$M_{offset10}$	0.132 (–0.332, 0.774)
M_{onset1}	**–0.778 (–1.469, –0.243**)
PVH–control	
Intercept	0.203 (–0.629, 1.114)
Age	–0.061 (–0.575, 0.388)
Sex (male)	0.078 (–0.081, 0.263)
$M_{offset10}$	–0.724 (–1.506, –0.181)
M_{onset1}	–0.791 (–1.559, –0.259)
Block 3	
NPVH–control	
Intercept	0.246 (–0.554, 1.182)
Age	–0.082 (–0.634, 0.459)
Sex (male)	0.042 (–0.113, 0.215)
$M_{offset10}$	–0.379 (–1.448, 0.436)
M_{onset1}	**–0.810 (–1.770, –0.036)**
STI_{offset}	**0.709 (0.033, 1.614)**
STI_{onset}	0.009 (–0.757, 0.715)
PVH–control	
Intercept	0.186 (–0.777, 1.176)
Age	–0.123 (–0.677, 0.361)
Sex (male)	0.062 (–0.116, 0.263)
$M_{offset10}$	**–1.374 (–2.920, –0.530)**
M_{onset1}	**–0.853 (–1.931, –0.097)**
STI_{offset}	**0.814 (0.090, 1.920)**
STI_{onset}	–0.002 (–1.006, 0.732)

Note. Bolded values indicate statistical significance (*p* < .05). NPVH = nonphonotraumatic vocal hyperfunction; PVH = phonotraumatic vocal hyperfunction; STI = spatiotemporal index.

Block 1 of the model—which included basic demographic parameters of age and sex (effect-coded as 0 = *female*, 1 = *male*)—did not improve model fit for discerning group compared to the null model with no predictors ($\chi^2_{\text{change}} = 3.57$, $p = .167$, pseudo-$R^2_{\text{adj}} = .04$). Mean RFF values entered in Block 2 led to statistically significant improvements in model fit ($\chi^2_{\text{change}} = 27.63$, $p < .001$) with pseudo-$R^2_{\text{adj}} = .59$. The introduction of STI values into the model in Block 3 also resulted in statistically significant improvements in fit ($\chi^2_{\text{change}} = 8.91$, $p = .012$) with pseudo-$R^2_{\text{adj}} = .68$. The coefficients from the final block of the hierarchical multinomial logistic regression model are presented in Table 3, and final equations from the model including STI are provided in Equations 1 and 2.

Model Evaluation: Training Data Set

One-vs-rest ROC analysis across bootstrapped samples was similar across all groups, with AUC values of 0.78 (95% CI [0.67, 0.87]) for the control group, 0.77 (95% CI [0.66, 0.87]) for the NPVH group, and 0.74 (95% CI [0.64, 0.83]) for the PVH group (see Table 4).

Three cutoff criteria were calculated by minimizing the Euclidean distance between sensitivity and specificity, to be used for subsequent model evaluation on the test data set. In the context of a multinomial logistic regression—with the control group set as the reference class—the classification process involves comparing the probabilities of each class against specific thresholds. Using the two final multinomial logistic regression equations (see Equations 1 and 2), probabilities for each class

Table 4. Receiver operating characteristic curve analysis on the training data set, shown as median (95% Wilson confidence interval) based on 10,000 bootstraps.

Model	Sensitivity	Specificity	Threshold	AUC
NPVH	0.73 (0.56, 0.89)	0.75 (0.58, 0.90)	0.35 (0.23, 0.49)	0.77 (0.66, 0.87)
PVH	0.73 (0.55, 0.90)	0.69 (0.52, 0.87)	0.34 (0.23, 0.47)	0.74 (0.64, 0.83)
Control	0.72 (0.55, 0.87)	0.78 (0.58, 0.92)	0.37 (0.25, 0.51)	0.78 (0.67, 0.87)

Note. AUC = area under the curve; NPVH = nonphonotraumatic vocal hyperfunction; PVH = phonotraumatic vocal hyperfunction.

(control, PVH, and NPVH) can be derived (summing to 1). These probabilities are then compared to the ROC-derived cutoff criteria (NPVH: 0.35, PVH: 0.34, control: 0.37; see Table 4). An instance is classified as the class whose probability exceeds its threshold and is the highest among the classes. For example, if the probabilities for a sample are *p*(NPVH) = .45, *p*(PVH) = .25, and *p*(Control) = .38, the instance would be classified as NPVH because .45 exceeds the NPVH threshold of .35 and demonstrates the highest probability. In cases where multiple classes exceed their thresholds or no class exceeds its threshold, the class with the highest probability is chosen.

Model Performance: Test Data Set

The regression model equations were applied to calculate predicted probabilities for participants in the untrained test data set. Using the thresholds obtained during model evaluation (see Table 4), group status predictions were performed by assigning each observation to the class with the highest probability exceeding its respective threshold.

One-vs-rest ROC curves were generated using the predictions from each test set (see Figure 3), yielding a median AUC of 0.78. For NPVH, the model achieved an AUC of 0.78 (sensitivity = 0.70, specificity = 0.75), consistent with an acceptable classifier (see Table 5). Classification performance was lowest for PVH, with an AUC of 0.69 (sensitivity = 0.60, specificity = 0.75). Despite lower PPVs of 0.58 for NPVH and 0.55 for PVH, the consistently high NPV of 0.75 across both groups suggests the model is more effective at ruling out VH than confirming it. For controls, the AUC was slightly higher at 0.86 (sensitivity = 0.70, specificity = 0.99), consistent with a strong classifier. The PPV and NPV of the control group were both 0.99, indicating excellent reliability in ruling out control cases.

A post hoc analysis revealed that, on average, misclassified samples exhibited higher overall severity of dysphonia ratings (M = 23.3, SD = 13.6, range: 4.6–48) compared to correctly classified samples (M = 15.3, SD = 12.7, range: 0–48). Misclassified samples also had fewer usable RFF traces per person at both voice offset (M = 8.8, SD = 3.7, range: 3–15) and onset (M = 8.1, SD = 3.5, range: 4–15) than correctly classified samples (offset: M = 10.5, SD = 5.0, range: 2–21; onset: M = 12.5, SD = 6.0, range: 2–27). No clear patterns in misclassifications were observed by recording site, age, or sex.

Discussion

Although RFF has shown promise as an acoustic measure for characterizing VH, it has not yet been widely adopted in clinical practice, partly because of the variability of mean RFF values observed within people. This study aimed to investigate this variability using two approaches: a basic measure of variability (standard deviation) and a more complex measure of variability that spans both magnitude and time (STI). Within this investigation, we examined acoustic voice signals that were recorded in a quiet room, providing an ecologically valid setting compared to the controlled conditions of a sound-attenuated booth commonly used in voice (and particularly, RFF) research.

Figure 3. One-versus-rest receiver operating characteristic curves, showing model performance in classifying group status of the test set (*n* = 30) as nonphonotraumatic vocal hyperfunction (NPVH), phonotraumatic vocal hyperfunction (PVH), or control.

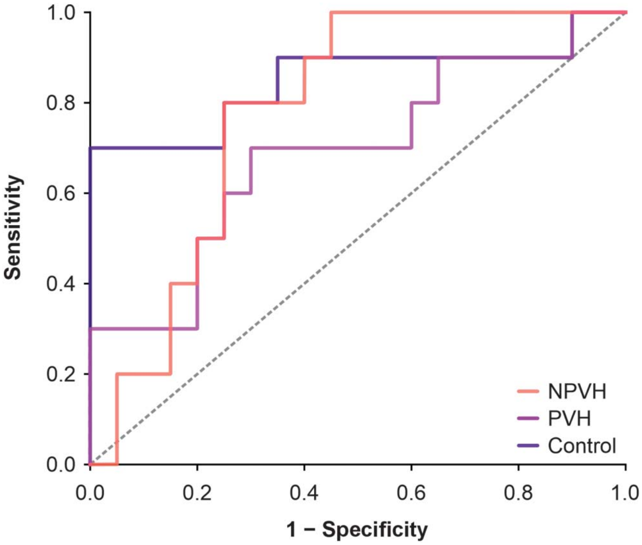


Table 5. Final regression model performance in predicting group membership on the test data set.

Model	AUC	Sensitivity	Specificity	PPV	NPV
NPVH	0.78	0.70	0.75	0.58	0.75
PVH	0.69	0.60	0.75	0.55	0.75
Control	0.86	0.70	0.99	0.99	0.99

Note. AUC = area under the curve; PPV = positive predictive value; NPV = negative predictive value; NPVH = nonphonotraumatic vocal hyperfunction; PVH = phonotraumatic vocal hyperfunction.

Consistent with past research, mean RFF significantly differed between individuals with and without VH (e.g., Ferrán et al., 2024; Roy et al., 2016; Stepp et al., 2010), even in this less controlled acoustic environment. The variability of RFF, via STI, significantly differed between individuals with and without VH. Furthermore, as hypothesized, incorporating STI revealed significant additional information for characterizing VH that was not captured by mean RFF alone. These findings highlight the importance of both RFF magnitude and variability in distinguishing VH and support the potential clinical utility of RFF measures in real-world settings.

Variability in Laryngeal Muscle Activity at Voice Offsets and Onsets

In the current study, we used two measures of variability to quantify the variability of RFF values across speech productions in VH. Our goal was to describe the consistency of how instantaneous voice *F*0 changes from a (pseudo)steady-state as individuals with VH stop and start voicing. We hypothesized that measures of variability would be higher in VH compared to age- and sex-matched controls and, further, that variability would be highest in the PVH subtype. These hypotheses stemmed from prior work suggesting that individuals with VH exhibit heightened baseline laryngeal tension, restricting their ability to modulate tension to start or stop voicing.

Overall, our results revealed complex patterns: Standard deviation demonstrated nonsignificant differences between VH subtypes and controls, whereas STI showed significant differences between these groups. The lack of group significance in standard deviation may be attributed to the different types of information considered by each measure. STI incorporates both magnitude (i.e., the size of the variations) and time information, whereas standard deviation only considers the magnitude of variability (i.e., the overall extent of variability without regard to whether it relates to size or time) of RFF values at voice offset (RFF_{offset10}) or onset (RFF_{onset1}). Specifically, STI captures the temporal pattern of variability across multiple cycles (Onsets 1–10 and Offsets 1–10), rather than focusing on single cycles (e.g., Onset 1 and Offset 10) as was done for standard deviation. Thus, STI may be more sensitive to subtle disruptions in sensorimotor control over time. We do not believe that statistical power alone explains the difference in performance, given that STI—though derived from more data points—is ultimately distilled into a single value. Rather, variability *across* the RFF trajectory appears to carry functional information, as supported by prior findings showing that patterns over cycles are relevant for both group discrimination (Kapsner-Smith et al., 2022) and perceptual severity modeling (Buckley et al., 2020). As such, STI is likely to be more sensitive to capturing complex variations in both magnitude and time across productions that may not be fully reflected by standard deviation, which only quantifies the magnitude of variability.

When considering our STI results, the PVH group exhibited the lowest variability relative to control and NPVH groups at voice offset; this difference was statistically significant only between the PVH and NPVH subtypes. This finding contrasts with our hypothesis that individuals with PVH would exhibit greater variability due to impaired control. One possible explanation, consistent with prior research, is that individuals with PVH may exhibit increased baseline laryngeal tension and reduced ability to modulate this tension (Hillman et al., 1989) potentially due to vocal fold lesions altering the physical properties of the vocal folds (e.g., mass, stiffness), leading to constrained voicing patterns (DeJonckere & Lebacq, 2022; Galindo et al., 2017; Jiang et al., 1998; Li et al., 2013). For instance, reduced variability may reflect compensatory strategies aimed at maintaining phonatory stability or could result from task simplification due to the nature of the stimuli. Because the current study infers sensorimotor function solely from acoustic output, we are limited in our ability to directly assess the physiological mechanisms underlying these group-level differences. As such, this unexpected finding is best viewed as hypothesis-generating rather than confirmatory. Future research incorporating additional methodologies—such as electromyography or imaging techniques—will be critical for determining whether reduced variability in PVH reflects lesion-induced rigidity, compensatory adaptations, or other contributing factors.

In support of our hypothesis, both VH subtypes showed elevated STI values at voice onset compared to controls. These findings are consistent with previous research examining voice onset time, which demonstrated that speakers with VH exhibit greater variability in phonemic voicing targets compared to controls (McKenna et al., 2020). Similar studies have observed higher voice onset time variability in individuals with PVH compared to controls (Marciniec, 2009) but not between VH subtypes (de Paula Soares et al., 2023; McKenna et al., 2020). Taken together, the observed variability at phonatory onset may reflect disrupted laryngeal motor control. However, since sensorimotor control was not directly assessed in this study, future research is needed to examine whether such mechanisms underlie the observed acoustic variability. One critical next step in this research is to therefore examine how specific sensorimotor disruptions give rise to these patterns and explore whether targeted interventions can reduce onset variability.

STI Captures Important Aspects of VH Not Reflected by Mean RFF Values Alone

We hypothesized that incorporating an estimate of RFF variability into a classification model that includes mean RFF values (RFF_{offset10}, RFF_{onset1}) would enhance the model's ability to discriminate between groups. As standard deviation did not show statistical significance relative to the effect of group, we focused our analysis on validating a classification model that incorporates the STI measure. To systematically evaluate the impact of introducing RFF measures for discerning group membership, we constructed a hierarchical multinomial logistic regression model and validated it using an independent test set. The baseline model included demographic information (age, sex), followed by the sequential addition of mean RFF measures and, subsequently, STI measures. Our results indicate that incorporating a complex measure of variability—STI—into the model significantly enhanced its ability to distinguish people with NPVH, PVH, and those without voice disorders, achieving acceptable discriminative performance as measured by AUC.

Prior work investigating the ability of RFF to distinguish groups primarily focused on one-on-one comparisons (e.g., NPVH vs. PVH) using RFF values alone (Heller Murray et al., 2017; Stepp et al., 2012) or when combined with other measures such as cepstral peak prominence (Kapsner-Smith et al., 2022). The results of this study underscore the significance of examining not just mean values but also the variability in hyperfunctional voice disorders. Our approach provides a more comprehensive representation of vocal function, potentially capturing subtle differences between groups that may be missed when relying solely on mean values. Moreover, the improved discriminative performance of the model that incorporates STI suggests that RFF variability may be a key factor in distinguishing between different types of VH and typical vocal function; this finding could have important downstream implications for clinical practice by highlighting within-person variability, potentially leading to more accurate diagnoses and tailored treatment plans for individuals with voice disorders.

The current study expands on foundational RFF research by addressing several methodological limitations found in prior studies. By collecting speech samples in a quiet environment and ensuring equal sample sizes across groups, we aimed to enhance the ecological validity and statistical robustness of our analysis, reducing the risk of Type I and Type II errors that can arise from unequal group representations (Rusticus & Lovato, 2014). This approach addresses potential limitations seen in prior studies, such as Heller Murray et al. (2017), in which unequal sample sizes may have impacted the generalizability of findings. Unlike previous studies that relied on sound-attenuated booths—which are not typically available in voice clinics—our approach captures speech in a more clinic-friendly setting to facilitate clinical accessibility. Our methods also incorporate a multinomial classification approach to provide a more holistic understanding of group differences. In doing so, we overcome the limitations of traditional pairwise comparisons using multiple binomial logistic regression analyses, which may obscure nuanced differences between groups due to potential correlations and asynchronous group likelihood estimates (Fitzhugh et al., 2023; Martin et al., 2021; Pate et al., 2023). By simultaneously examining how each group differs from the collective others, we can better understand the distinctive characteristics of PVH and NPVH. Overall, the methodological decisions of this study increase the generalizability of our findings in understanding the complex biomechanical mechanisms underlying different VH presentations.

The novelty of using STI to examine RFF opens new avenues for understanding vocal function in VH. Traditional approaches have primarily focused on mean RFF values, which may not fully capture the complexity of laryngeal motor control and the dynamic nature of voice production. This limitation may contribute to the observed variability in RFF values within and across individuals with and without VH. By incorporating STI measures along with mean values, researchers may gain deeper insights into the patterns of variability in vocal function.

While STI is traditionally used in articulatory kinematics to assess trial-to-trial variability in movement trajectories, we have implemented it here to assess temporal variability in RFF traces. We do not mean to imply a direct parallel between RFF-based STI measures and articulatory kinematics—where STI is traditionally used to assess trial-to-trial variability in movement trajectories—but rather acknowledge that the temporal patterning (or variability) in RFF traces may provide meaningful insights into vocal function. We recognize that this adaptation of STI is a novel and exploratory approach, and we acknowledge the limitations of inferring underlying motor variability or pathology from STI in the absence of articulatory or physiological data. This method is subject to further validation, and we have revised the article to clarify this point explicitly.

The use of STI in voice acoustics holds potential for enhancing the precision of voice assessments by identifying subtle differences in laryngeal motor control that may not be apparent through traditional mean analysis alone. In this study, we focused on STI for RFF. However, other estimates of STI—such as those for voice fundamental frequency, sound pressure level, or cepstral peak prominence, where standard deviation already holds clinical significance (Patel et al., 2018)—may also contain valuable information about vocal function. Incorporating STI measures into clinical voice evaluations could aid in the discrimination of VH subtypes and improve diagnostic accuracy.

Limitations and Future Directions

It is important to acknowledge the inherent limitations of RFF: It can be influenced by several factors, including irregular vocal fold vibration (e.g., aperiodicity, vocal fry), external variables such as recording environment and speaker-specific features such as overall dysphonia severity (Vojtech et al., 2019). Our post hoc findings underscore the relevance of these factors, as misclassified samples tended to have higher overall dysphonia severity ratings and fewer usable RFF traces on average; these characteristics may have contributed to noisier or less stable acoustic estimates, potentially reducing classification accuracy. Thus, while the number of usable repetitions did not show a systematic effect on the initial outcome measures (M_{offset10}, M_{onset1}, SD_{offset10}, SD_{onset1}, STI_{offset}, STI_{onset}; see Supplemental Material S1), our final results suggest that the number usable traces may play a more significant role in downstream model performance than previously appreciated. Careful consideration of these factors is therefore essential to ensure that STI-based analyses of RFF yield meaningful and clinically relevant insights into vocal function in VH.

While STI appears to offer complementary value to mean RFF in differentiating groups, the interpretation of STI as a marker of motor variability remains inferential. Validation using physiological correlates—such as subglottal pressure variation, electromyography, or articulatory kinematics—is essential to confirm its role as an indicator of neuromotor control in VH. Importantly, this study does not aim to directly measure laryngeal physiology or neural control mechanisms. We instead focus on group-level acoustic patterns that may reflect underlying differences in phonatory behavior. As such, the approaches described here may enhance our ability to characterize acoustic patterns in VH and inform hypotheses about underlying mechanisms of VH. Future research should explore STI estimates across multiple vocal parameters to establish comprehensive norms and further elucidate the intricacies of laryngeal motor control in both clinical and nonclinical populations. In addition, follow-up studies using multivariate discriminant approaches with larger data sets should be conducted to help define clear classification boundaries and improve clinical interpretability of STI-based models.

Although STI presents a promising approach for assessing vocal function, there are several practical limitations to its implementation in clinical settings. Currently, the calculation of STI requires multiple repetitions of RFF tasks along with custom processing steps, including resampling and normalization procedures. These steps introduce computational complexity that is not readily available in typical clinical software systems. Additionally, the need for multiple valid traces for accurate STI calculation may limit its feasibility in fast-paced clinical environments, where time constraints often prevail. To make STI more accessible in clinical settings, targeted software development is needed to automate and streamline these processes. As a first step, our MATLAB-based script for computing STI has been uploaded for open-source use.[5] Future work should therefore focus on the development of user-friendly tools that guide data collection, processing, and STI calculation with minimal clinician input. Until these tools are developed, implementing STI in clinical practice will remain limited by the current computational and data collection requirements.

Conclusions

Traditional methods of calculating RFF have relied on mean values to characterize laryngeal tension in hyperfunctional voice disorders. This study demonstrates that incorporating RFF variability in both time and magnitude—via STI—provides valuable, statistically significant information about the presence and type of VH, offering insights beyond what mean values yield alone. These findings suggest that STI may uncover subtle differences in laryngeal motor control that may be overlooked by traditional RFF analyses. Future research on RFF should consider this variability when establishing norms and assessing treatments over time. Exploring the utility of STI across various acoustic parameters may additionally enhance our understanding of VH and improve the precision of clinical voice evaluations.

[5] Available from https://sites.bu.edu/stepplab/research/rff/.

Author Contributions

Jenny Vojtech: Conceptualization, Investigation, Methodology, Data curation, Formal analysis, Software, Validation, Visualization, Writing – original draft. **Laura E. Toles:** Conceptualization, Data curation, Writing – review & editing. **Daniel P. Buckley:** Data curation, Writing – review & editing. **Cara E. Stepp:** Conceptualization, Investigation, Methodology, Project administration, Resources, Writing – review & editing, Supervision, Funding acquisition.

Data Availability Statement

The data sets analyzed during the current study are not publicly available due to identifiable information.

Acknowledgments

The authors would like to thank Aakash Patel and Harsha Haridas for their help preparing audio samples for analysis. This research was supported by R01 DC015570 (Cara E. Stepp) from the National Institute on Deafness and Other Communication Disorders.

References

Abur, D., Subaciute, A., Kapsner-Smith, M., Segina, R. K., Tracy, L. F., Noordzij, J. P., & Stepp, C. E. (2021). Impaired auditory discrimination and auditory-motor integration in hyperfunctional voice disorders. *Scientific Reports, 11*(1), Article 13123. https://doi.org/10.1038/s41598-021-92250-8

Altman, K. W., Atkinson, C., & Lazarus, C. (2005). Current and emerging concepts in muscle tension dysphonia: A 30-month review. *Journal of Voice, 19*(2), 261–267. https://doi.org/10.1016/j.jvoice.2004.03.007

American Speech-Language-Hearing Association. (2002). *Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) form.* https://www.asha.org/form/cape-v/

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology, 26*(1), 32–46. https://doi.org/10.1111/j.1442-9993.2001.01070.pp.x

Anderson, M. J. (2017). Permutational multivariate analysis of variance (PERMANOVA). In *Wiley StatsRef: Statistics reference online* (pp. 1–15). Wiley. https://doi.org/10.1002/9781118445112.stat07841

Aronson, A. E., & Bless, D. M. (2009). *Clinical voice disorders* (4th ed.). Thieme Medical.

Askenfelt, A. G., & Hammarberg, B. (1986). Speech waveform perturbation analysis: A perceptual-acoustical comparison of seven measures. *Journal of Speech and Hearing Research, 29*(1), 50–64. https://doi.org/10.1044/jshr.2901.50

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

Belsky, M. A., Rothenberger, S. D., Gillespie, A. I., & Gartner-Schmidt, J. L. (2021). Do phonatory aerodynamic and acoustic measures in connected speech differ between vocally healthy adults and patients diagnosed with muscle tension dysphonia? *Journal of Voice, 35*(4), 663.e1–663.e7. https://doi.org/10.1016/j.jvoice.2019.12.019

Bhattacharyya, N. (2014). The prevalence of voice problems among adults in the United States. *The Laryngoscope, 124*(10), 2359–2362. https://doi.org/10.1002/lary.24740

Buckley, D. P., Diaz Cadiz, M., Eadie, T. L., & Stepp, C. E. (2020). Acoustic model of perceived overall severity of dysphonia in adductor-type laryngeal dystonia. *Journal of Speech, Language, and Hearing Research, 63*(8), 2713–2722. https://doi.org/10.1044/2020_JSLHR-19-00354

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

Colton, R. H., Casper, J. K., & Leonard, R. (2006). *Understanding voice problems: A physiological perspective for diagnosis and treatment*. Lippincott Williams & Wilkins.

Crocker, C., Toles, L. E., Morrison, R. A., & Shembel, A. C. (2024). Relationships between vocal fold adduction patterns, vocal acoustic quality, and vocal effort in individuals with and without hyperfunctional voice disorders. *Journal of Voice*. Advance online publication. https://doi.org/10.1016/j.jvoice.2023.12.012

DeJonckere, P. H., & Lebacq, J. (2022). Vocal fold collision speed in vivo: The effect of loudness. *Journal of Voice, 36*(5), 608–621. https://doi.org/10.1016/j.jvoice.2020.08.025

Dejonckere, P. H., Remacle, M., Fresnel-Elbaz, E., Woisard, V., Crevier-Buchman, L., & Millet, B. (1996). Differentiated perceptual evaluation of pathological voice quality: Reliability and correlations with acoustic measurements. *Revue de Laryngologie-Otologie-Rhinologie, 117*(3), 219–224.

de Paula Soares, M. F., Sampaio, M., & Brockmann-Bauser, M. (2023). Interaction of voice onset time with vocal hyperfunction and voice quality. *Applied Sciences, 13*(15), Article 8956. https://doi.org/10.3390/app13158956

Dworkin, J. P., Meleca, R. J., & Abkarian, G. G. (2000). Muscle tension dysphonia. *Current Opinion in Otolaryngology & Head and Neck Surgery, 8*(3), 169–173. https://doi.org/10.1097/00020840-200006000-00007, 173.

Dworkin-Valenti, J. P., Stachler, R. J., Stern, N., & Amjad, E. H. (2018). Pathophysiologic perspectives on muscle tension dysphonia. *Archives of Otolaryngology and Rhinology, 4*(1), 001–10. https://doi.org/10.17352/2455-1759.000065

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. https://doi.org/10.1044/2017_JSLHR-S-16-0337

Ferrán, S., Rodríguez-Zanetti, C., Garaycochea, O., Terrasa, D., Prieto-Matos, C., Del Río, B., Alzuguren, M. P., & Fernández, S. (2024). Relative fundamental frequency: Only for hyperfunctional voices? A pilot study. *Bioengineering, 11*(5), Article 475. https://doi.org/10.3390/bioengineering11050475

Fitzhugh, N., Rasmussen, L. R., Simoni, A. H., & Valentin, J. B. (2023). Misuse of multinomial logistic regression in stroke related health research: A systematic review of methodology. *European Journal of Neuroscience, 58*(4), 3116–3131. https://doi.org/10.1111/ejn.16084

Free, N., Stemple, J. C., Smith, J. A., & Phyland, D. J. (2023). Variability in voice characteristics of female speakers with phonotraumatic vocal fold lesions. *Journal of Voice*. Advance online publication. https://doi.org/10.1016/j.jvoice.2023.01.019

Galindo, G. E., Peterson, S. D., Erath, B. D., Hillman, R. E., & Zañartu, M. (2017). Modeling the pathophysiology of phonotraumatic vocal hyperfunction with a triangular glottal model of the vocal folds. *Journal of Speech, Language, and Hearing Research, 60*(9), 2452–2471. https://doi.org/10.1044/2017_JSLHR-S-16-0412

Gillespie, A. I., Gartner-Schmidt, J., Rubinstein, E. N., & Abbott, K. V. (2013). Aerodynamic profiles of women with muscle tension dysphonia/aphonia. *Journal of Speech, Language, and Hearing Research, 56*(2), 481–488. https://doi.org/10.1044/1092-4388(2012/11-0217)

Goberman, A. E., & 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

Groll, M. D., Peterson, S. D., Zañartu, M., Vojtech, J. M., & Stepp, C. E. (2022). Empirical evaluation of the role of vocal fold collision on relative fundamental frequency in voicing offset. *Journal of Voice, 39*(2), 345–352. https://doi.org/10.1016/j.jvoice.2022.09.016

Hanschmann, H., Lohmann, A., & Berger, R. (2010). Comparison of subjective assessment of voice disorders and objective voice measurement. *Folia Phoniatrica et Logopaedica, 63*(2), 83–87. https://doi.org/10.1159/000316140

Heller Murray, E. S., Lien, Y. S., Van Stan, J. H., Mehta, D. D., Hillman, R. E., Pieter Noordzij, J., & Stepp, C. E. (2017). Relative fundamental frequency distinguishes between phonotraumatic and non-phonotraumatic vocal hyperfunction. *Journal of Speech, Language, and Hearing Research, 60*(6), 1507–1515. https://doi.org/10.1044/2016_JSLHR-S-16-0262

Higgins, M. B., Chait, D. H., & Schulte, L. (1999). Phonatory air flow characteristics of adductor spasmodic dysphonia and muscle tension dysphonia. *Journal of Speech, Language, and Hearing Research, 42*(1), 101–111. https://doi.org/10.1044/jslhr.4201.101

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

Hillman, R. E., Stepp, C. E., Van Stan, J. H., Zañartu, M., & Mehta, D. D. (2020). An updated theoretical framework for vocal hyperfunction. *American Journal of Speech-Language Pathology, 29*(4), 2254–2260. https://doi.org/10.1044/2020_AJSLP-20-00104

Hogikyan, N. D., & Sethuraman, G. (1999). Validation of an instrument to measure voice-related quality of life (V-RQOL). *Journal of Voice, 13*(4), 557–569. https://doi.org/10.1016/s0892-1997(99)80010-1

Jacobson, B. H., Johnson, A., Grywalski, C., Silbergleit, A., Jacobson, G., Benninger, M. S., & Newman, C. W. (1997). The Voice Handicap Index (VHI): Development and validation. *American Journal of Speech-Language Pathology, 6*(3), 66–70. https://doi.org/10.1044/1058-0360.0603.66

Jafari, N., Salehi, A., Izadi, F., Talebian Moghadam, S., Ebadi, A., Dabirmoghadam, P., Faham, M., & Shahbazi, M. (2017). Vocal function exercises for muscle tension dysphonia: Auditory-perceptual evaluation and self-assessment rating. *Journal of Voice, 31*(4), 506.e25–506.e2531. https://doi.org/10.1016/j.jvoice.2016.10.009

Jiang, J. J., Diaz, C. E., & Hanson, D. G. (1998). Finite element modeling of vocal fold vibration in normal phonation and hyperfunctional dysphonia: Implications for the pathogenesis of vocal nodules. *Annals of Otology, Rhinology & Laryngology, 107*(7), 603–610. https://doi.org/10.1177/000348949810700711

Jiang, J. J., & Titze, I. R. (1994). Measurement of vocal fold intraglottal pressure and impact stress. *Journal of Voice, 8*(2), 132–144. https://doi.org/10.1016/S0892-1997(05)80305-4

Kapsner-Smith, M. R., Díaz-Cádiz, M. E., Vojtech, J. M., Buckley, D. P., Mehta, D. D., Hillman, R. E., Tracy, L. F., Noordzij, J. P., Eadie, T. L., & Stepp, C. E. (2022). Clinical cutoff scores for acoustic indices of vocal hyperfunction that combine relative fundamental frequency and cepstral peak prominence. *Journal of Speech, Language, and Hearing Research, 65*(4), 1349–1369. https://doi.org/10.1044/2021_jslhr-21-00466

Koufman, J. A., & Blalock, P. D. (1991). Functional voice disorders. *Otolaryngologic Clinics of North America, 24*(5), 1059–1073. https://doi.org/10.1016/S0030-6665(20)31068-9

Li, Z., Bakhshaee, H., Helou, L., Mongeau, L., Kost, K., Rosen, C., & Verdolini, K. (2013). Evaluation of contact pressure in human vocal folds during phonation using high-speed videoendoscopy, electroglottography, and magnetic resonance imaging. *Proceedings of Meetings on Acoustics, 19*(1), Article 060306. https://doi.org/10.1121/1.4800732

Lien, Y. S., 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. https://doi.org/doi:10.1044/2014_JSLHR-S-13-0158

Löfqvist, A., Baer, T., McGarr, N. S., & Story, R. S. (1989). The cricothyroid muscle in voicing control. *The Journal of the Acoustical Society of America, 85*(3), 1314–1321. https://doi.org/10.1121/1.397462

Lowell, S. Y., Kelley, R. T., Colton, R. H., Smith, P. B., & Portnoy, J. E. (2012). Position of the hyoid and larynx in people with muscle tension dysphonia. *The Laryngoscope, 122*(2), 370–377. https://doi.org/10.1002/lary.22482

Marciniec, S. A. (2009). *Voice onset time of women with vocal nodules*. Rush University.

Martin, G. P., Sperrin, M., Snell, K. I. E., Buchan, I., & Riley, R. D. (2021). Clinical prediction models to predict the risk of multiple binary outcomes: A comparison of approaches. *Statistics in Medicine, 40*(2), 498–517. https://doi.org/10.1002/sim.8787

Maryn, Y., Dick, C., Vandenbruaene, C., Vauterin, T., & Jacobs, T. (2009). Spectral, cepstral, and multivariate exploration of tracheoesophageal voice quality in continuous speech and sustained vowels. *The Laryngoscope, 119*(12), 2384–2394. https://doi.org/10.1002/lary.20620

Mathieson, L., Hirani, S. P., Epstein, R., Baken, R. J., Wood, G., & Rubin, J. S. (2009). Laryngeal manual therapy: A preliminary study to examine its treatment effects in the management of muscle tension dysphonia. *Journal of Voice, 23*(3), 353–366. https://doi.org/10.1016/j.jvoice.2007.10.002

McArdle, B. H., & Anderson, M. J. (2001). Fitting multivariate models to community data: A comment on distance-based redundancy analysis. *Ecology, 82*(1), 290–297. https://doi.org/10.1890/0012-9658(2001)082[0290:FMMTCD]2.0.CO;2

McDowell, S., Morrison, R., Mau, T., & Shembel, A. C. (2022). Clinical characteristics and effects of vocal demands in occupational voice users with and without primary muscle tension dysphonia. *Journal of Voice, 39*(2), 448–456. https://doi.org/10.1016/j.jvoice.2022.10.005

McKenna, V. S., Hylkema, J. A., Tardif, M. C., & Stepp, C. E. (2020). Voice onset time in individuals with hyperfunctional voice disorders: Evidence for disordered vocal motor control. *Journal of Speech, Language, and Hearing Research, 63*(2), 405–420. https://doi.org/10.1044/2019_JSLHR-19-00135

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

Morrison, M. D., Rammage, L. A., Belisle, G. M., Pullan, C. B., & Nichol, H. (1983). Muscular tension dysphonia. *Journal of Otolaryngology, 12*(5), 302–306.

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika, 78*(3), 691–692. https://doi.org/10.1093/biomet/78.3.691

Nanjundeswaran, C., Jacobson, B. H., Gartner-Schmidt, J., & Verdolini Abbott, K. (2015). Vocal Fatigue Index (VFI): Development and validation. *Journal of Voice, 29*(4), 433–440. https://doi.org/10.1016/j.jvoice.2014.09.012

Pate, A., Riley, R. D., Collins, G. S., van Smeden, M., Van Calster, B., Ensor, J., & Martin, G. P. (2023). Minimum sample size for developing a multivariable prediction model using multinomial logistic regression. *Statistical Methods in Medical Research, 32*(3), 555–571. https://doi.org/10.1177/09622802231151220

Patel, R. R., Awan, S. N., Barkmeier-Kraemer, J., Courey, M., Deliyski, D., Eadie, T., Paul, D., Svec, J. G., Hillman, R., 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. https://doi.org/10.1044/2018_ajslp-17-0009

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*(85), 2825–2830. https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf [PDF]

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

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

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

Rusticus, S. A., & Lovato, C. Y. (2014). Impact of sample size and variability on the power and Type I error rates of equivalence tests: A simulation study. *Practical Assessment, Research & Evaluation, 19*(11). https://eric.ed.gov/?id=EJ1038106

Shipurkar, R., White, M., & Katz, W. (2023). Normalization of speech kinematic data for characterizing primary muscle tension dysphonia. *The Journal of the Acoustical Society of America*, *154,* Article A337. https://doi.org/10.1121/10.0023721

Smith, A., Johnson, M., McGillem, C., & Goffman, L. (2000). On the assessment of stability and patterning of speech movements. *Journal of Speech, Language, and Hearing Research, 43*(1), 277–286. https://doi.org/10.1044/jslhr.4301.277

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

Stepp, C. E., Lester-Smith, R. A., Abur, D., Daliri, A., Pieter Noordzij, J., & Lupiani, A. A. (2017). Evidence for auditory-motor impairment in individuals with hyperfunctional voice disorders. *Journal of Speech, Language, and Hearing Research, 60*(6), 1545–1550. https://doi.org/10.1044/2017_JSLHR-S-16-0282

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

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

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

Thomas, C. M., Rhodes, D., Mehta, M., & Alexander, J. (2023). Methods of measuring laryngeal muscle tension in patients with muscle tension dysphonia: A scoping review. *Journal of Voice*. Advance online publication. https://doi.org/10.1016/j.jvoice.2023.03.013

van den Berg, J. (1958). Myoelastic–aerodynamic theory of voice production. *Journal of Speech and Hearing Research, 1*(3), 227–244. https://doi.org/10.1044/jshr.0103.227

Van Houtte, E., Van Lierde, K., D'Haeseleer, E., & Claeys, S. (2010). The prevalence of laryngeal pathology in a treatment-seeking population with dysphonia. *The Laryngoscope, 120*(2), 306–312. https://doi.org/10.1002/LARY.20696

van Mersbergen, M. R., Beckham, B. H., & Hunter, E. J. (2021). Do we need a measure of vocal effort? Clinician's report of vocal effort in voice patients. *Perspectives of the ASHA Special Interest Groups, 6*(1), 69–79. https://doi.org/10.1044/2020_PERSP-20-00258

Vojtech, J. M., Segina, R. K., Buckley, D. P., Kolin, K. R., Tardif, M. C., Noordzij, J. P., & Stepp, C. E. (2019). Refining algorithmic estimation of relative fundamental frequency: Accounting for sample characteristics and fundamental frequency estimation method. *The Journal of the Acoustical Society of America, 146*(5), 3184–3184. https://doi.org/10.1121/1.5131025

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. https://doi.org/10.1121/1.423068

Wisler, A., Goffman, L., Zhang, L., & Wang, J. (2022). Influences of methodological decisions on assessing the spatiotemporal stability of speech movement sequences. *Journal of Speech, Language, and Hearing Research, 65*(2), 538–554. https://doi.org/10.1044/2021_JSLHR-21-00298

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