Refining algorithmic estimation of relative fundamental frequency: Accounting for sample characteristics and fundamental frequency estimation method

Jennifer M. Vojtech[a)]
Department of Biomedical Engineering, Boston University, 44 Cummington Mall, Boston, Massachusetts 02215, USA

Roxanne K. Segina
Department of Speech, Language, and Hearing Sciences, Boston University, 635 Commonwealth Avenue, Boston, Massachusetts 02215, USA

Daniel P. Buckley[a)]
Department of Otolaryngology—Head and Neck Surgery, Boston University School of Medicine, 72 East Concord Street, Boston, Massachusetts 02118, USA

Katharine R. Kolin and Monique C. Tardif
Department of Speech, Language, and Hearing Sciences, Boston University, 635 Commonwealth Avenue, Boston, Massachusetts 02215, USA

J. Pieter Noordzij
Department of Otolaryngology—Head and Neck Surgery, Boston University School of Medicine, 72 East Concord Street, Boston, Massachusetts 02118, USA

Cara E. Stepp[b),c)]
Department of Speech, Language, and Hearing Sciences, Boston University, 635 Commonwealth Avenue, Boston, Massachusetts 02215, USA

(Received 30 May 2019; revised 7 October 2019; accepted 8 October 2019; published online 11 November 2019)

Relative fundamental frequency (RFF) is a promising acoustic measure for evaluating voice disorders. Yet, the accuracy of the current RFF algorithm varies across a broad range of vocal signals. The authors investigated how fundamental frequency (f_o) estimation and sample characteristics impact the relationship between manual and semi-automated RFF estimates. Acoustic recordings were collected from 227 individuals with and 256 individuals without voice disorders. Common f_o estimation techniques were compared to the autocorrelation method currently implemented in the RFF algorithm. Pitch strength-based categories were constructed using a training set (1158 samples), and algorithm thresholds were tuned to each category. RFF was then computed on an independent test set (291 samples) using category-specific thresholds and compared against manual RFF via mean bias error (MBE) and root-mean-square error (RMSE). Auditory-SWIPE′ for f_o estimation led to the greatest correspondence with manual RFF and was implemented in concert with category-specific thresholds. Refining f_o estimation and accounting for sample characteristics led to increased correspondence with manual RFF [MBE = 0.01 semitones (ST), RMSE = 0.28 ST] compared to the unmodified algorithm (MBE = 0.90 ST, RMSE = 0.34 ST), reducing the MBE and RMSE of semi-automated RFF estimates by 88.4% and 17.3%, respectively.
https://doi.org/10.1121/1.5131025

[AL] Pages: 3184–3202

I. INTRODUCTION

Approximately one-third of adults in the United States report suffering from a voice problem during their lifetime (Bainbridge *et al.*, 2017; Bhattacharyya, 2014; Roy *et al.*, 2005). Nearly a quarter of these individuals report recurrent issues (Roy *et al.*, 2005). Despite the prevalence of voice problems, current voice assessments are primarily subjective, including interpretations of patient history, psychosocial questionnaires, auditory-perceptual assessments, and physical evaluations (Morrison *et al.*, 1986; Roy *et al.*, 2013; Schwartz *et al.*, 2009). Although these methods provide some insight into vocal health, the reliability of such techniques is variable among raters (Poburka *et al.*, 2017; Stepp *et al.*, 2011a; Yiu *et al.*, 2014; Zraick *et al.*, 2011), potentially leading to inconsistencies in interpretation. As such, investigations have turned to identifying objective measures

[a)] Also at: Department of Speech, Language, and Hearing Sciences, Boston University, 635 Commonwealth Avenue, Boston, MA 02215, USA. Electronic mail: jmvo@bu.edu

[b)] Also at: Department of Otolaryngology—Head and Neck Surgery, Boston University School of Medicine, Boston, MA 02215, USA.

[c)] Also at: Department of Biomedical Engineering, Boston University, 44 Cummington Mall, Boston, MA 02215, USA.

that may be effective adjuncts to current subjective methods for voice assessment (Dejonckere *et al.*, 2001; Maryn and Weenink, 2015).

Vocal strain, defined as the perception of tenseness or excessive vocal effort associated with phonation (Hirano, 1981; Kempster *et al.*, 2009), is a prominent feature of clinical voice quality assessments. Strain is thought to be associated with excessive or imbalanced laryngeal muscle forces (Askenfelt and Hammarberg, 1986; Lowell *et al.*, 2012), which in turn have been implicated in a variety of voice disorders, including vocal hyperfunction (Hillman *et al.*, 1989; Morrison *et al.*, 1986), spasmodic dysphonia (Ludlow, 2009; Roy *et al.*, 1996), and Parkinson's disease (Gallena *et al.*, 2001). Of the auditory-perceptual features assessed in clinical voice quality assessments, however, vocal strain is the least reliable feature (Dejonckere *et al.*, 1996). Research efforts have turned toward identifying an objective means of assessing vocal strain; given the non-invasive nature of acquiring acoustic signals, it may be a promising modality for developing an objective estimator of vocal strain. Although abnormal levels of laryngeal muscle tension are thought to be related to a dysphonic voice quality, an acoustic indicator specific to vocal strain does not exist (Bhuta *et al.*, 2004; Mehta and Hillman, 2008). For instance, vocal strain was shown to strongly correlate with cepstral peak prominence, sharpness, and various spectral moments (Anand *et al.*, 2018); yet, this study was limited in sample size. In recent studies, relative fundamental frequency (RFF) has also received attention as a promising acoustic measure for specifically assessing (Eadie and Stepp, 2013; Lien *et al.*, 2015b) and tracking (Stepp *et al.*, 2011b) this perception of excessive vocal effort.

A. Current investigation

Quantitative measures of vocal strain are needed to improve assessment and track clinical progress. Since RFF has shown promise as an acoustic estimator of vocal strain, we aimed to build upon the previous work of Lien *et al.* (2017) to establish methodology for estimating RFF across a wide range of vocal signals. To improve the semi-automated RFF estimation algorithm, we compared an assortment of fundamental frequency estimation algorithms to the method currently employed in the semi-automated RFF algorithm, and used a training set of speech samples to tune algorithmic parameters for computing RFF based on individual sample attributes. Finally, the error between manual and algorithmic RFF values was evaluated in an independent test set. This work aims to improve the potential clinical applicability of RFF related to estimating vocal strain for use in conjunction current clinical voice assessment techniques.

B. Relative fundamental frequency

1. Overview

During a voiced sonorant–voiceless consonant–voiced sonorant (VCV) production, RFF captures the instantaneous changes in fundamental frequency (f_o) corresponding to the transition into and out of the voiceless consonant. These changes in f_o are dependent on the vibration of the vocal folds, which is, in turn, a function of vocal fold length, mass, and tension (Van Den Berg, 1958). RFF is computed from the ten instantaneous f_o values before ("offset cycles") and after ("onset cycles") the voiceless consonant. These instantaneous f_o values are normalized to the steady state f_o of the nearest voiced sonorant in a VCV production (see Fig. 1). Changes in RFF during these transitions form a characteristic pattern in speakers with healthy voices, which has been attributed to interactions of laryngeal muscle tension, vocal fold kinematics, and changes in airflow (Löfqvist *et al.*, 1989; Stepp *et al.*, 2011b; Stevens, 1977; Van Den Berg, 1958; Watson, 1998). There is evidence to support that laryngeal muscle tension is transiently elevated before, during, and after the production of the voiceless consonant in order to inhibit vocal fold vibration (Löfqvist *et al.*, 1989; Stevens, 1977). As glottal abduction commonly occurs for voiceless sounds, it is postulated that vocal fold abductory kinematics act in concert with elevated muscle tension to achieve devoicing during the transition into the voiceless consonant (Watson, 1998). The transition out of the voiceless consonant is hypothesized to occur as a result of the interplay between increases in laryngeal muscle tension and airflow from the preceding sonorant (Watson, 1998), in addition to vocal fold adductory kinematics necessary to bring the vocal folds together and reinitiate vibration.

With the interactions among these three physiological mechanisms in mind, recent work has shown that RFF is capable of differentiating between healthy and disordered voices characterized by excessive laryngeal tension, such as

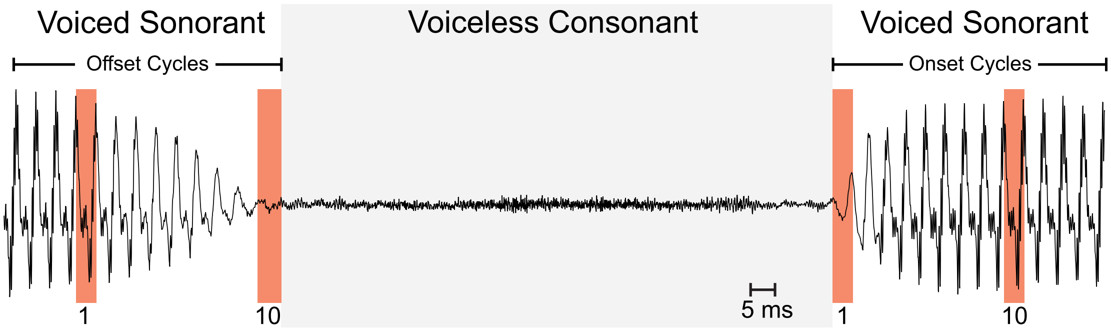

FIG. 1. (Color online) Acoustic waveform of a voiced sonorant–voiceless consonant–voiced sonorant (VCV) instance, with voice offset and voice onset cycles identified. The first and tenth cycles of each voiced sonorant are highlighted. Voice offset cycles are normalized to offset cycle 1, whereas voice onset cycles are normalized to voice onset cycle 10.

vocal hyperfunction (Stepp *et al.*, 2010; Stepp *et al.*, 2011b), adductor spasmodic dysphonia (Eadie and Stepp, 2013), and Parkinson's disease (Goberman and Blomgren, 2008; Stepp, 2013). Excessive laryngeal tension is thought to be associated with increased vocal effort, and increased vocal effort may result in a strained voice quality in some speakers (McKenna and Stepp, 2018). Vocal strain also covaries with the perceptual qualities of breathiness and roughness in approximately 50% of speakers (Lowell *et al.*, 2012). It then follows that RFF has also been shown to correlate with auditory-perceptual judgments of dysphonia severity (Roy *et al.*, 2016; Stepp *et al.*, 2012), which encompasses multiple dimensions of voice quality—including breathiness, roughness, and strain—and to quantify the degree of laryngeal tension. Specifically, Lien *et al.* (2015b) found that when healthy individuals modulated their vocal effort, they achieved RFF values that were not only different from their typical patterns but were similar to those observed in individuals with tension-based voice disorders. As such, RFF could be a useful tool to not only identify differences in laryngeal tension between individuals but to track changes in laryngeal tension within an individual over time.

2. Semi-automated RFF estimation

Although RFF shows promise as a quantitative measure of laryngeal muscle tension, calculation via manual extraction is tedious and time-consuming. This is because a trained technician must visually examine the acoustic waveform during voice offset or onset transitions, then exercise trial-and-error to identify the vocal cycle marking the termination or initiation of voicing, respectively. One commonly used software for this is Praat (Boersma, 2001). With this software, the technician must identify and use the location of this boundary cycle (i.e., voice offset cycle 10, voice onset cycle 1) to further identify the 11 glottal pulse timings that correspond to the edges of the 10 vocal cycles closest to the voiceless consonant. Selecting the boundary cycle and ensuring that the adjacent glottal pulse timings are accurately estimated by Praat is a tedious task since f_o estimation is particularly difficult near voice offsets (voiced-to-unvoiced transitions) and onsets (unvoiced-to-voiced transitions; Quatieri, 2008). Not only can vocal cycle masking occur due to environmental noise or concurrent aspiration and frication from coarticulation, but instantaneous f_o estimation implies stationarity. Thus, calculating f_o during the transition from voiced to unvoiced speech is challenging. Additionally, it has been reported that at least six RFF speech sequences are needed for a reliable RFF estimate, wherein extensively trained technicians must perform 20–40 min of analysis per reliable RFF extraction (Eadie and Stepp, 2013; Lien *et al.*, 2017). Despite being the current gold-standard technique, manual RFF extraction is clinically infeasible due to its time- and training-intensive nature. Thus, in order to minimize the need for inefficient, manual intervention by trained technicians, semi-automated RFF estimation has been developed using rule-based signal processing techniques (Lien, 2015; Lien *et al.*, 2017).

a. Current RFF algorithm design. Current semi-automated algorithm[1] (henceforth referred to as the "V.1" algorithm, which stands for version 1 algorithm) estimates RFF in six steps: (1) identification of the fricatives and vowels in each utterance according to the acoustic signal via high-to-low energy ratios; (2) estimation of average f_o via autocorrelation of the vowels; (3) identification of peaks and troughs of potential vocal cycles pertaining to the voiced sonorant; (4) identification of boundaries between each voiced sonorant and the voiceless consonant via the normalized peak-to-peak amplitude, number of zero crossings, and waveform shape similarity; (5) rejection of instances that did not meet specified criteria (e.g., less than ten onset or offset cycles, glottalization, misarticulation, voiceless consonant is voiced); and (6) RFF calculation. RFF is computed as follows (Lien, 2015; Lien *et al.*, 2017):

$$\mathrm{RFF}\,(\mathrm{ST}) = 12 \times \log_2\left(\frac{f_o}{f_o^{\mathrm{ref}}}\right). \tag{1}$$

RFF (semitones, ST) is calculated by comparing the instantaneous fundamental frequency of a vocal cycle (f_o) to the approximate steady-state fundamental frequency (f_0^{ref}), which was considered as the f_o of the first offset cycle for offset RFF values and the tenth onset cycle for onset RFF values. This formula for calculating RFF is used in both manual and automated RFF estimation. Within this algorithm, all steps are fully automated except for step (1), wherein the location of the fricatives may require manual intervention. Of the six steps in the V.1 algorithm design, step (4) is especially prone to error due to the complexity of f_o estimation at voice offsets and onsets (Quatieri, 2008).

b. Issues with the semi-automated RFF algorithm.

1. Method of f_o estimation

The second and third steps of the semi-automated algorithm utilize f_o estimation in order to identify potential vocal cycles in the voiced sonorant of a VCV production. The acoustic signal of a VCV utterance is first bandpass filtered according to the average f_o of the speaker. Then, a sliding window is constructed using the inverse of the average f_o of the speaker; this sliding window shifts from the identified midpoint of the fricative in the first step and moves toward the voiced sonorant of interest. Potential vocal cycles are identified by leveraging the normalized peak-to-peak amplitude, number of zero crossings, and waveform shape similarity obtained from each sliding window.

One issue that plagues the V.1 algorithm is a reliance on autocorrelation-based f_o estimation to calculate the mean f_o of each vowel in a VCV production. Typical f_o estimation methods operate either by comparing a segment of the speech signal with other segments that have been offset by a certain period or examining the frequency content of the signal. In the former case, multiple windows of time are compared to identify a repeating pattern (e.g., discontinuities, peak, or trough amplitudes) in the waveform that may provide insight into potential values of f_o; in the latter case, the signal is transformed from the time domain to the frequency domain, wherein energy peaks occur at integer multiples of the f_o. The majority of f_o estimation techniques generate pitch period candidates based on assumptions that the rate of vocal fold vibration generally varies a small percentage from one period to the next, and the configuration of the vocal tract varies at a rate much slower than the rate of the vocal folds (Talkin, 1995). Autocorrelation was originally selected for the RFF algorithm because Praat, the acoustic analysis software used to perform manual RFF calculations, makes use of autocorrelation for f_o estimation. Although autocorrelation-based f_o estimation provides high resolution in the time domain with relatively low computational complexity, autocorrelation assumes signal periodicity. While vocally healthy speech signals generally present with a prominent peak corresponding to the speaker's pitch period, the level of aperiodicity in a speech signal increases when voice problems are present (Eadie and Doyle, 2005; Titze, 1995). As such, performing simple autocorrelation analyses on voices characterized as dysphonic may lead to increases in measurement error. Additionally, even healthy voices are described as *quasiperiodic* at best. In many cases, multiple autocorrelation peaks are present, and one problem with implementing autocorrelation is determining which peak correctly corresponds to the pitch period (Rabiner, 1977). Most importantly, autocorrelation requires 2–3 complete pitch periods in order to examine physiological f_o ranges encountered in speech. As such, rapid changes in f_o captured by RFF (e.g., during the transition into and out of a vowel during a VCV production) may lead to estimation inaccuracies and poor cycle-to-cycle resolution.

The shortcomings surrounding the use of autocorrelation for f_o estimation during voiced/unvoiced speech transitions may be overcome in different f_o detection methods. For instance, normalized cross correlation is an alternate time-domain method that compares frames of the original signal to subsampled frames of the signal (Talkin, 1995). Normalized cross correlation operates on smaller segments of the speech signal and, thus, may be useful for examining f_o during such voiced/unvoiced transitions. Modifications may also be performed on the basic autocorrelation function (e.g., weighting harmonics, scoring pitch period candidates) in order to minimize its flaws regarding signal periodicity and poor cycle-to-cycle resolution. Therefore, an investigation into different f_o estimation methods is warranted in order to determine which f_o estimation method results in the greatest correspondence between algorithmic and manual RFF estimates.

2. Sample characteristics

(a) Motivation

Lien *et al.* (2017) sought to validate the V.1 algorithm as a faster, more objective means of computing RFF, thus, expanding on its potential for use in clinical voice assessment. In the study, the V.1 algorithm was evaluated against manual RFF estimation in a heterogenous group of voice samples that varied in recording location and/or equipment, diagnosis, voice quality, and/or dysphonia severity. The authors included a group of typical voice speakers ($N = 36$) to serve as controls to a larger group of individuals with disordered voices ($N = 154$). Approximately half of the speakers were recorded in a waiting area or quiet room of a voice clinic with signals sampled at 44.1 kHz and 16-bit resolution, whereas the remaining speakers were recorded in a sound-treated room with signals sampled at 20 kHz and 16-bit resolution. This dataset was further split into a training set of 126 speakers and 2 testing sets composed of the remaining 64 speakers. Although the V.1-derived RFF estimates were highly correlated with manual RFF estimates, the authors noted that the relationship between RFF estimation methods was dependent on voice sample characteristics, specifically noting dysphonia severity and recording location as potential factors influencing this relationship. Indeed, speech signals recorded from a waiting area or quiet room of a voice clinic resulted in a poorer correlation (0.82 versus 0.91) and root-mean-squared error (0.37 ST versus 0.28 ST) between V.1 and manual RFF estimates than those recorded in a sound-treated room; however, they were, on average, also more dysphonic. Therefore, the authors could not evaluate whether dysphonia severity, recording location, or a mix of these two characteristics led to the observed variations in V.1 performance. One future direction of this work, thus, included evaluating V.1 algorithm performance across the spectrums of dysphonia severity and signal acquisition quality.

With this background in mind, it is important to consider how the widely ranging characteristics of the voice samples may have affected V.1 performance. As previously mentioned, the fourth step of the V.1 algorithm leverages a set of acoustic features to identify the boundary between voiced and voiceless segments. As a sliding window moves from the voiceless consonant into the voiced sonorant, the normalized peak-to-peak amplitude, number of zero crossings, and waveform shape similarity between adjacent cycles are extracted from each window. Within the examined window, normalized peak-to-peak amplitude is calculated as the range of the speech signal, number of zero crossings is computed as the number of sign changes of the speech signal, and waveform shape similarity is calculated as the normalized sum of square error between the current window of time and the previous window of time. If a positive or negative peak is identified in the region of the voiceless consonant, the normalized peak-to-peak amplitude is expected to be low, number of zero crossings to be high, and waveform shape similarity to be high (Lien, 2015). The V.1 algorithm uses the acoustic feature vectors to locate the transition point between the two speech segments, called the "boundary cycle." It is expected that the largest change in the three acoustic feature vectors will occur at this transition; as such, the V.1 algorithm identifies the vector index that maximizes the effect size between the left and right components of each acoustic feature vector. A single set of thresholds is applied to each acoustic feature vector to assist in identifying the boundary cycle for voice offset (offset cycle 10) and onset (onset cycle 1); a positive or negative peak that satisfies at least two out of the three thresholds is chosen as the boundary cycle for that offset or onset instance. In the V.1 algorithm, these thresholds were optimized across a heterogenous group of voice samples to minimize the overall difference between manual and semi-automated RFF estimates (Lien *et al.*, 2017). Specifically, the threshold for each acoustic feature was adjusted in step sizes of 5% from 80% to 120%; the step that resulted in the best performance in the training set was chosen as the threshold for that acoustic feature.

Since a single threshold set is used to identify voice offset and onset boundary cycles, boundary cycle identification varies based on the voice samples used to train and test the V.1 algorithm. This is because the threshold set—which was optimized across a heterogenous set of voice samples—may not be the best set for individual voice samples. For instance, a healthy voice sample may be more periodic than many of the samples used to train the semi-automated algorithm to calculate RFF; in this scenario, the thresholds that identify the boundary cycle are tuned to more aperiodic signals, but the criteria for choosing the boundary cycle may differ for more periodic signals. Complications in identifying the boundary cycle across a wide range of speech signals stems from vocal cycle masking and a lack of f_o stationarity at voice offsets and onsets. Manual RFF estimation is not as impacted by these complications since trained technicians are able to visualize the acoustic waveform and subjectively choose the boundary cycle, employing trial-and-error techniques when necessary. Because differences in voice sample characteristics have been shown to affect the V.1 algorithm performance, it is necessary to take these differences into consideration prior to RFF computation. This is a critical hurdle to overcome prior to transferring RFF to clinical use.

(b) Accounting for sample characteristics

One method of accounting for differences in voice sample characteristics is to develop categories that enable RFF estimation via category-specific thresholds instead of using a single set of thresholds for all voice samples. Thus, instead of using the same acoustic feature thresholds to identify the boundary cycles across all samples, it is possible that evaluating a voice sample according to its specific attributes (i.e., categorizing it) and applying a set of acoustic thresholds that are optimal for the voice sample will increase correspondence between V.1 and manual RFF estimates. For the purposes of this study, two acoustic speech sample characteristics were hypothesized to be important to algorithmic accuracy: overall severity of dysphonia and signal acquisition quality. Overall severity of dysphonia is an auditory-perceptual measure included within the Consensus-Auditory Perceptual Evaluation of Voice (CAPE-V). Overall dysphonia severity is considered the global, integrated impression of vocal deviance (Kempster *et al.*, 2009), whereas signal acquisition quality encompasses features of the signal acquisition and the room conditions in which a speech sample is recorded, such as the recording environment. As part of transferring RFF for use in clinical voice assessments, the algorithm should be able to reliably calculate RFF, regardless of the signal acquisition quality; since clinicians may not have access to sound-treated rooms, we examine RFF values from acoustic recordings collected not only in sound-treated rooms, but also in the waiting area or a quiet room of a voice clinic.

A single acoustic measure, pitch strength, may be sensitive to both overall severity of dysphonia (Eddins *et al.*, 2016; Kopf *et al.*, 2017; Shrivastav *et al.*, 2012) and signal acquisition quality, and as such, may be used to objectively quantify these sample characteristics. Pitch strength describes the saliency of pitch sensation and can be calculated using the auditory sawtooth waveform inspired pitch estimator—Prime (Auditory-SWIPE′) model (Camacho, 2012). In Auditory-SWIPE′, the f_o of an acoustic signal is estimated using a multi-step process, described here in brief. An input sound is first filtered using an auditory-processing front end. The spectrum of the sound is then compared to a sawtooth wave across a range of f_o values; to minimize subharmonic errors, only the fundamental and prime harmonics of the signal are considered. Each sawtooth wave is correlated to the spectrum of the filtered input sound; the f_o of the sawtooth wave that elicits the highest degree of correlation is labeled as the pitch of the input sound. The level of correlation (from 0 to 1) is labeled as the estimated pitch strength of the input sound. Thus, sounds with higher pitch sensations result in higher pitch strengths, whereas sounds with lower pitch sensations result in lower pitch strengths.

Pitch strength has been implemented in the objective assessment of voice quality due to its versatility across voice signal types[2] (Kopf *et al.*, 2017). For instance, a perceptually breathy speech signal may be classified as containing some level of stochastic noise due to the turbulence surrounding the airflow jet when the voice is produced. Despite lacking an obvious f_o, the signal may still elicit a pitch sensation, and therefore, a non-zero pitch strength. Indeed, pitch strength has been shown to be correlated with perceptual judgments of voice quality (Eddins *et al.*, 2016; Shrivastav *et al.*, 2012) and recently proposed as a treatment outcome for dysphonia (Kopf *et al.*, 2017). Because of this, pitch strength may be a viable, objective measure that can assess overall severity of dysphonia.

Pitch strength may also be of interest to assess signal acquisition quality. Numerous factors may impact the signal acquisition quality of a speech signal, including speaker characteristics (e.g., distance, loudness) and recording environment. As such, the degree of room reverberation, degree of room noise, and proximity of the acquisition to reflecting surfaces must be taken into consideration when capturing acoustic speech signals (Titze, 1995); moreover, environmental noise can significantly affect acoustic correlates of voice quality (Deliyski *et al.*, 2005). With this in mind, introducing noise into a speech signal—such as environmental noise or the noisy by-product of a turbulent airstream generated at the glottis—will reduce the harmonics-to-noise ratio of the signal (Awan and Frenkel, 1994), which may, in turn, impact the pitch strength of the signal. Thus, pitch strength may be a useful measure to encompass signal acquisition quality as it pertains to the speaker being recorded and the environment in which the speaker is recorded.

II. METHODS

A. Participants

Informed consent was obtained, and the study carried out in compliance with the Boston University (BU) and the University of Washington (UW) Institutional Review Boards. A group of 256 adults without voice disorders (152 female, 104 male)[3] aged 18–100 yr [mean $(M) = 37.6$ yr, standard deviation (SD) = 22.3 yr] participated in the study, all of whom reported no prior history of speech, language, or hearing disorders. A group of 227 individuals with disordered voices (148 female, 79 male)[3] aged 18–84 yr ($M = 52.9$ yr, $\mathrm{SD} = 17.7$ yr) also participated in the study. All participants with Parkinson's disease were diagnosed with idiopathic Parkinson's disease by a neurologist, and were recorded while on their typical carbidopa/levodopa medication schedule. Individuals with deep brain stimulation devices were requested to switch their devices off for the duration of the study. All other participants within this group were diagnosed with a voice disorder by a board-certified laryngologist (see Table VI in the Appendix for diagnosis frequency of participants with disordered voices).

A speech-language pathologist specializing in voice disorders assessed the overall severity of dysphonia[4] (0–100) of each participant using the CAPE-V (Kempster *et al.*, 2009). Table I shows participant demographics, separated according to group, sex, age, and overall dysphonia severity. The speech-language pathologist reanalyzed 15% of samples in a separate sitting to ensure adequate intrarater reliability. The Pearson's product-moment correlation coefficient for CAPE-V ratings conducted on 15% of reassessed speech samples was calculated using the statistical package R (R Core Team, 2013; version 3.2.4), yielding $r = 0.96$.

B. Recording procedures

1. Equipment and environment

All microphone signals were digitally recorded and analysis occurred offline. Participants were recorded in either (a) a waiting area or quiet room at Boston Medical Center using a dynamic headset microphone (model WH20XLR; Shure, Niles, IL) or at BU using a condenser headset microphone (model SM35XLR; Shure, Niles, IL), (b) a sound-treated room at BU using the same condenser headset microphone, sampling rate, and resolution, or (c) a quiet room at UW using a dynamic headset microphone (model WH20XLR; Shure, Niles, IL). All microphone signals were sampled at 44.1 kHz with 16-bit resolution. The directional headset microphone was placed 45° from the midline and 7–10 cm from the lips. Recording levels were not standardized across speakers so as to determine the impact of signal acquisition quality on RFF values.

2. Speaker training

Participants were instructed to produce three sets of nonsense words at their typical pitch and loudness of conversational speech. Each set of nonsense words contained three repetitions of a VCV instance involving the voiceless consonant /f/. Specifically, participants were instructed to produce three /afa/ repetitions, take a breath, produce three /ifi/ repetitions, take a breath, and produce three /ufu/ repetitions (i.e., /afa afa afa/, breath, /ifi ifi ifi/, breath, /ufu ufu ufu/). VCV productions were recorded rather than running speech to minimize intraspeaker variability (Lien *et al.*, 2014). Similarly, the token /f/ was used to minimize individual variation within speaker (Lien *et al.*, 2014). In order to maintain typical pitch and loudness, participants were instructed to refrain from chanting or singing the production. Additionally, participants were instructed to repeat the utterance if any of the three repetitions were misarticulated or glottalized.

Three speech samples—each comprising three VCV productions—were collected from each of the 483 speakers; this resulted in 4347 VCV instances within 1449 speech samples. Figure 2 shows an overview of this data breakdown, as well as those for the training and test sets used throughout Sec. II C 2 to refine the semi-automated RFF algorithm.

C. Data analysis

Recordings from each participant were entered into a database of RFF stimuli for further analysis. This RFF database contained samples across the spectrum of signal quality and dysphonia severity to ensure a full representation of the range of diversity present in clinical practice. This database includes samples recorded in worst-case conditions (i.e., low signal acquisition quality and high dysphonia severity) and best-case conditions (i.e., high signal acquisition quality and low dysphonia severity). A larger proportion of female voices were entered into this database due to the higher prevalence of voice disorders in adult females than in adult males (Martins *et al.*, 2016; Roy *et al.*, 2005). Figure 3 shows the relative frequency of overall dysphonia severity ratings with respect to the recording environment, described as a quiet room or waiting area versus a sound-attenuated room. Of the 483 speakers, 335 (207 speakers without voice disorders, 128 speakers with voice disorders) were recorded in a sound-treated room and 148 (49 speakers without voice disorders, 99 speakers with voice disorders) were recorded in a quiet room or waiting area. The purpose of including a large variety of sample characteristics was to ensure that RFF estimates would not be confounded by specific signal quality and dysphonia severity attributes in order to minimize the effect of the populations used to train and test the algorithm on resulting algorithmic performance.

1. Manual RFF estimation

All recordings were manually analyzed using Praat acoustic analysis software (Boersma, 2001) via methodology

TABLE I. Summary of participant demographics.

	Sex (*N*)		Age (years)			Overall severity of dysphonia		
Group	F	M	Mean	Standard deviation	Range	Mean	Standard deviation	Range
Speakers without disordered voices	152	104	37.6	22.3	18–100	11.5	8.1	0–44.6
Speakers with disordered voice	148	79	52.9	17.7	18–84	22.1	20.0	0–100

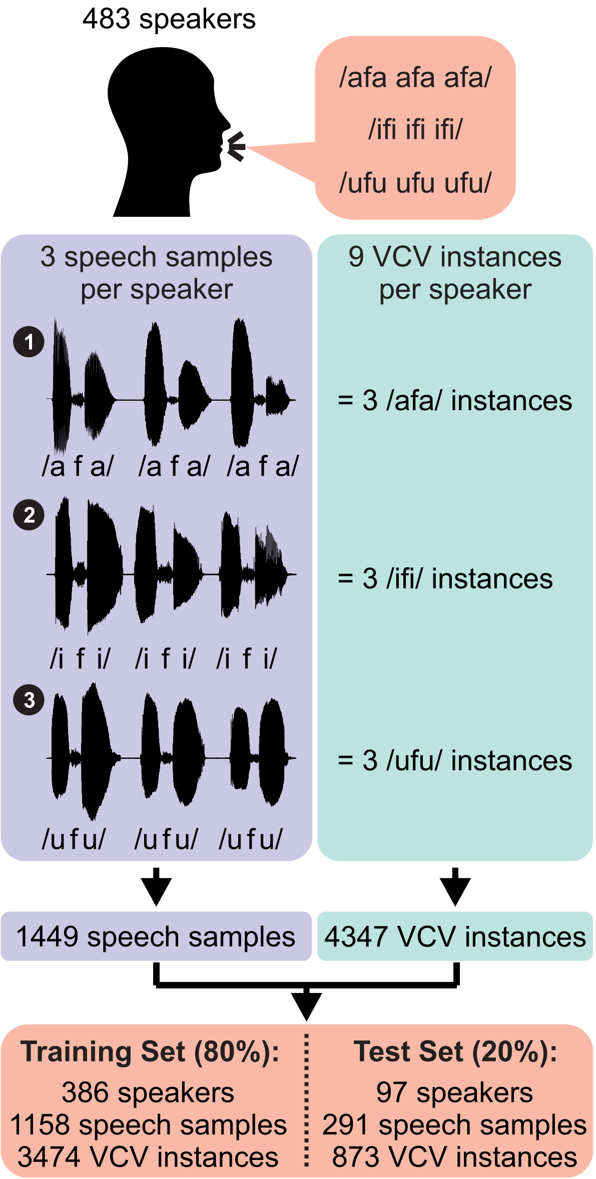

FIG. 2. (Color online) Speech sample collection flowchart, wherein speakers produced three different VCV utterances (i.e., /afa/, /ifi/, /ufu/) three times each (i.e., /afa afa afa/, /ifi ifi ifi/, /ufu ufu ufu/). This led to a total of nine VCV productions collected from each of 483 speakers. Speakers were split into training (80%) and test (20%) sets via simple random sampling.

previously described in Vojtech and Heller Murray (2019). For each speech sample, the f_o analysis range in Praat was initially set to 60–300 Hz for male recordings and 90–500 Hz for female recordings; these settings were manually adjusted for individuals whose pitch range fell outside of the default pitch range. Manual RFF estimation proceeded as follows: (i) a trained technician determined the suitability of each of the RFF instances, wherein instances were excluded if a phoneme was misarticulated, if glottalization was observed in either voiced portion of the sample, or if either of the voiced sections failed to reach steady-state (Lien *et al.*, 2015a; Lien and Stepp, 2013), (ii) the technician identified the ten adjacent offset and onset cycles nearest to the voiceless phoneme /f/ (refer to Fig. 1), (iii) the instantaneous f_o was calculated as the inverse of the difference between adjacent pulse periods, and (iv) RFF (in ST) was calculated according to Eq. (1) (Baken and Orlikoff, 2000).

Although a single technician has been shown to have high internal reliability when manually extracting RFF values (Lien *et al.*, 2015a; Lien *et al.*, 2015b), the ability to distinguish the vocal cycle closest to the fricative (i.e., offset cycle 10 and onset cycle 1) can vary between technicians and, ultimately, impact RFF cycle values. To minimize the possibility of such variation, a minimum of two technicians carried out pulse period selection and RFF computation for each sample. For manual RFF ratings, each of 9 VCV productions from 483 speakers were rated by at least 2 trained technicians. Due to the availability of technicians to perform manual RFF estimates, eight trained technicians manually calculated RFF throughout the course of data collection. Three trained technicians (6–8) completed this training and performed manual RFF estimates prior to 2015, whereas the remaining trained technicians (1–5) did so after 2015. All raters were trained[5] in interrater reliability > 0.93. Table II displays the number of speakers that each of eight trained technicians rated, including the number of speaker RFF estimates that overlapped with other raters. Of the 483 speakers, 437 were rated by 2 trained technicians (technicians 1–8 in Table II) and 46 were rated by 3 trained technicians (technicians 6–8 in Table II).

Average RFF values were computed across technicians as the gold-standard for RFF estimation. Intrarater and interrater reliability metrics were calculated using the statistical package R (version 3.2.4). Interrater reliability was conducted on the RFF estimates using a two-way random intraclass correlation (ICC[2,k]) to evaluate the precision of these manual, gold-standard estimates of RFF. Intrarater reliability was then computed via Pearson correlation coefficients within each technician when asked to re-estimate 15% of their samples in a different sitting (Lien *et al.*, 2017; Lien *et al.*, 2015b). The average intrarater reliability was calculated as $r = 0.92$ (SD $= 0.04$, range $= 0.87$–0.99). Average interrater reliability of RFF estimates was computed as ICC $= 0.92$ (SD $= 0.05$, range $= 0.82$–0.99).

2. Semi-automated RFF Estimation

Building off of the algorithm developed by Lien (2015), analyses were performed in MATLAB (version 9.3; MathWorks, Natick, MA) in order to evaluate the effect of the f_o estimation method and sample characteristics on resulting RFF values.

a. Method of f_o estimation. Fundamental frequency estimation is used in two major steps of the V.1 algorithm (see Sec. I B 2 a). First, the f_o estimation method is used to compute the average f_o of the speaker. The average f_o of the speaker is used to create a sliding window that moves along the acoustic waveform to collect positive and negative peaks, in addition to extracting three acoustic features from the speech segment contained within the window. Following, f_o estimation is used to calculate the pitch period between peaks identified during the sliding window process; these peaks are referred to as "vocal cycle candidates." As the boundary

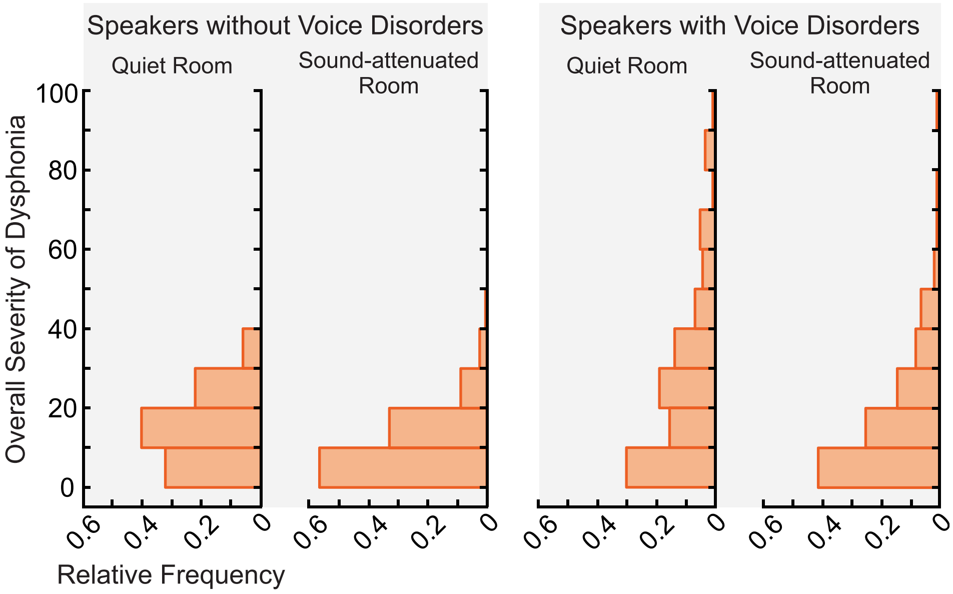

FIG. 3. (Color online) Relative frequency of overall severity of dysphonia ratings for speakers with (left) and without (right) voice disorders, further distinguished by whether the speaker was recorded in a quiet room or sound-attenuated room.

cycle is identified from the vocal cycle candidates by examining the set of acoustic features extracted along the sliding window, these steps are paramount to accurate RFF estimation.

Two scenarios were selected to evaluate the correspondence between manual and automated RFF estimates with respect to the two steps in that algorithm that rely on accurate f_o estimation. In the first scenario, the midpoint of the voiceless consonant was provided to the V.1 algorithm to evaluate the ability of the f_o estimation method to generate a sliding window for capturing potential peaks and troughs, then to calculate the pitch period of vocal cycle candidates; in other words, both steps of vocal cycle detection were examined in this scenario. In the second scenario, the manually defined indices of the boundary cycle were provided in order to solely evaluate the ability of the f_o estimation method to calculate the pitch period of *actual* vocal cycles used in RFF computation; thus, the ability of the f_o estimation method to determine vocal cycle candidates was not evaluated.

In order to determine the impact of the f_o estimation method on the resulting RFF values, we compared five f_o detection methods to manual RFF values. In addition to the autocorrelation function that is currently implemented in V.1, four competitive f_o estimators were chosen for evaluation: Auditory-SWIPE′ (Camacho, 2012), YIN (de Cheveigne and Kawahara, 2002), robust algorithm for pitch tracking (RAPT; Talkin, 1995), and Halcyon (Azarov*et al.*, 2016). To generate potential pitch period candidates, Auditory-SWIPE′ and YIN use modified autocorrelation function models, whereas RAPT and Halcyon each employ normalized cross-correlation function-based models. These methods were selected due to their superior performance and/or frequency of use as f_o estimators (Azarov *et al.*, 2016; Jouvet and Laprie, 2017).

To compare the f_o estimation techniques, RFF was first calculated on a subset of 180 speech samples (9 speech samples from 20 participants) in each of the aforementioned scenarios: (1) when the approximate midpoint of the voiceless consonant was provided, and (2) when the manually defined indices of the boundary cycle were provided. A small subset of speech samples was chosen for this analysis to model a spectrum of pitch strength values. Of the subset of 20 participants, 15 speakers were recorded in a sound-attenuated room, whereas the remaining 5 were recorded in a quiet room or waiting area of a voice clinic. Moreover, 11 of the 20 individuals were diagnosed with a voice disorder. This provided pitch strength values ranging from 0.04 to 0.51 ($M = 0.33$, $SD = 0.12$). Error metrics were calculated between manual and semi-automated RFF estimates via mean bias error (MBE), root-mean-square error (RMSE), and the number of erroneous rejections. RMSE and MBE were included to provide an overview of the accuracy and precision of automated RFF values, respectively, when compared to ground truth manual RFF estimates. Erroneous rejections were considered as RFF instances that were excluded only through automated analysis (false positive) or manual analysis (false negative). A false positive occurred when an instance did not meet specified criteria according to the V.1 algorithm (see Lien, 2015, for these criteria) but was considered valid through manual analysis; conversely, a false negative occurred when an instance was considered misarticulated, glottalized, or not steady-state through

TABLE II. Number of speakers for which eight trained technicians manually computed relative fundamental frequency. The matrix shows common speakers analyzed between technicians, whereas the diagonal (bold) describes the number of speakers a single technician rated in total.

Technician	1	2	3	4	5	6	7	8
1	**103**							
2	93	**278**						
3	1	92	**188**					
4	0	79	9	**91**				
5	9	0	86	3	**99**			
6	0	0	0	0	0	**96**		
7	0	1	0	0	1	95	**97**	
8	0	13	0	0	0	47	46	**60**

manual analysis (Lien *et al.*, 2015a; Lien and Stepp, 2013), but was considered valid through automated analysis. The number of erroneous rejections was tallied for each of the two aforementioned scenarios as the sum of false positives and false negatives. Resulting MBE, RMSE, and erroneous rejections were computed when augmenting the V.1 algorithm with each of the five f_o estimation techniques. In comparing error metrics, the f_o estimation method that led to the smallest error (i.e., RMSE, MBE) and least number of erroneous rejections was retained as the f_o estimation method in a new algorithm version, termed "V.2."

b. Development of category-specific thresholds. The RFF database consisted of 1449 speech samples (4347 VCV instances) from 483 independent speakers. This database was split up into training (80%) and test (20%) sets using simple random sampling, as this distribution method leads to low bias of model performance (Kuhn and Johnson, 2013; Reitermanova, 2010); since simple random sampling can lead to high variance in model performance, k-fold cross-validation was used to quantify this variation (see Sec. II C 2 b 4). Pitch strength—which was adopted in the current study to quantify sample characteristics (i.e., speaker voice quality and signal acquisition quality)—was then calculated for each VCV instance in the training set via Auditory-SWIPE′. Following, the training set was used to tune the boundary between voiced and voiceless segments in VCV instances according to sample characteristics.

1. Automated sample rejection

First, a rejection threshold was created to eliminate samples with a pitch strength considered too low (i.e., little-to-no presence of a pitch sensation) to accurately analyze. The rejection cutoff was determined by constructing a receiver operating characteristic (ROC) curve using pitch strength estimates of instances that were rejected via manual analysis versus those considered valid. A final threshold was chosen using the threshold obtained at the location in the ROC curve corresponding to the maximum positive likelihood ratio (PLR); the maximum PLR was chosen as criterion for the rejection cutoff to maximize the probability of rejecting a sample that is invalid, and minimize the probability of rejecting a viable sample. Any VCV instances with a pitch strength below this cutoff value were excluded from further analysis.

2. Boundary cycle shifts

Prior to category creation, we evaluated the average discrepancy in boundary cycle identification between manual and semi-automated RFF analyses. As previously mentioned, the methods of determining the boundary between voiced and voiceless speech segments differs between manual and automated RFF analysis: manual analysis allows technicians to subjectively choose the boundary between voiced and voiceless speech segments, whereas the V.1 algorithm leverage a set of acoustic features to identify this transition point. In the V.1 algorithm, Lien *et al.* (2017) extracted acoustic features during the sliding window process (i.e., acoustic feature values collected at each average pitch period, starting at the voiceless consonant and transitioning into the voiced sonorant) with the assumption that each feature would exhibit a state transition at the boundary between voiced and voiceless speech segments. This logic was implemented by maximizing the effect size of each acoustic feature vector, wherein either side of the state transition was assumed to contain stable values pertaining to the voiced sonorant and voiceless consonant. The index that maximized the vector effect size was considered the boundary cycle for that acoustic feature, and the median boundary cycle candidate of the three acoustic features was adopted as the "true" boundary cycle.

In order to evaluate the differences in manual versus automated boundary extraction, we first excluded any of the 4347 VCV instances (from 1158 speech samples) of the training set that were rejected through manual analysis. We then compared the "automated" boundary cycle that was computed via the V.1 algorithm to the true boundary cycle, which was determined via manual analysis. First, the acoustic features implemented in the V.1 algorithm were examined as a function of the average number of pitch periods away from the true boundary cycle. This included normalized peak-to-peak amplitude, number of zero crossings, and waveform shape similarity; however, a preliminary analysis showed that bandpass filtering the microphone signal 3 ST above and below the average f_o of the sample (Lien, 2015) increased the effect size for normalized peak-to-peak amplitude. As such, normalized peak-to-peak amplitude was computed from the filtered microphone signal for all analyses. After calculating the three acoustic features as a function of pitch periods away from the true boundary, the effect sizes of each feature at each distance ±2 pitch periods from the true boundary cycle were computed (see Table VII in the Appendix). Following the logic for selecting boundary cycles by Lien *et al.* (2017), the pitch period cycle that elicited the maximum effect size was characterized as the automated boundary cycle for that acoustic feature. This automated boundary cycle—represented as the average number of pitch periods away from the true boundary cycle—corresponded to the error in boundary selection between algorithmic and manual RFF estimation. The acoustic feature slope at the pitch period cycle was characterized as the direction of the error (i.e., toward or away from the voiceless consonant).

3. Category creation

VCV instances with average pitch strength values that did not exceed the rejection threshold discussed in Sec. II C 2 b 1 were excluded from further analysis. Pitch strength estimates of the remaining VCV instances were then used to construct categories for computing RFF according to a sample's attributes. For all samples in the training set, acoustic feature values calculated at the time point of the automated boundary cycle were evaluated against pitch strength. Each feature vector was then visually inspected across pitch strength, paying particular attention to local extrema, inflection points, and confidence intervals; this elicited pitch strength categories for voice offset and voice onset. Specifically, categories were chosen by leveraging these variables to identify pitch strength levels that were represented by consistent increases, decreases, or stable feature values.

ROC curves were then constructed for each offset and onset feature by pitch strength category in order to determine the discriminatory ability of pitch strength to objectively distinguish between features at the true versus automated boundary. Optimal thresholds were chosen to distinguish categories via the Youden index (Youden, 1950) in order to maximize sensitivity and specificity. Using this logic, the automated boundary cycle would be determined via maximizing the effect size of the acoustic feature vector per sample as in V.1. However, the acoustic feature value at the selected boundary cycle would then undergo an additional analysis: if the value of the acoustic feature at the boundary cycle did not exceed the threshold set by the sample's pitch strength category, then the automated boundary cycle would be shifted to mark the correct transition between voiced and voiceless segments. The number of vocal cycles and direction of the shift corresponded to those determined in Sec. II C 2 b 2 (see Table VII in the Appendix). Notably, however, the V.1 algorithm collects positive and negative waveform peaks from the speech signal. RFF is then computed from the set of peaks that is closer in time to the voiceless consonant. Because of this, an additional analysis was performed to further refine RFF computation via the pitch strength-tuned categories. Specifically, the median distance between automated and true boundary cycles was computed for each voice offset and onset category when identifying positive and negative waveform peaks. The difference in average pitch periods between the true and automated boundary cycles was then implemented as a final shift to increase correspondence with manual RFF boundary cycle identification.

4. Validation and performance

RFF was recalculated in the training set using the optimal set of acoustic feature thresholds for each sample category. Then, *k*-fold cross-validation was performed on the recalculated RFF values of the training data to provide an appropriate estimate of model performance and ensure that the model was not overfitted (Kuhn and Johnson, 2013). For this analysis, the training dataset was split into *k*-training (1042 speech samples containing 3126 VCV instances) and *k*-validation (116 speech samples containing 343 VCV instances) datasets, then RMSE and MBE values were compared as an average across $k = 10$ folds.

The semi-automated RFF algorithm that incorporated optimized f_o estimation (V.2) and accounted for sample characteristics via pitch strength categories were termed "V.3." RFF estimation accuracy was then computed on the test set by comparing manual RFF estimates against V.1–V.3. Error was computed across the three algorithm versions via RMSE and MBE to provide insight into the average accuracy and precision of each algorithm, respectively. Additionally, the impact of clinical sample characteristics (i.e., dysphonia severity and signal acquisition quality) on the V.3 algorithm was evaluated using the test set. Two Welch's tests were conducted for resulting MBE and RMSE values in order to investigate the performance of the V.3 algorithm in computing RFF on samples of differing signal acquisition qualities. Average errors for samples recorded in a quiet room/waiting area versus a sound-attenuated room were examined using an alpha level of 0.05 for significance testing. Effect sizes were estimated using Cohen's *d*. Then, the performance of the V.3 algorithm in estimating RFF across the spectrum of dysphonia severity was evaluated by calculating Pearson product-moment correlation coefficients for MBE and RMSE values against overall severity of dysphonia ratings.

III. RESULTS

A. Method of f_o estimation

The RMSE, MBE, and a number of erroneous rejections for each f_o estimation technique are shown in Table III for $N = 20$ speakers. When the manually determined boundary cycle was provided as a reference, the RFF algorithm using Auditory-SWIPE′ (RMSE = 0.52 ST) and Halcyon (MBE = 0.03 ST) for f_o estimation resulted in the greatest correspondence to manual RFF. Notably, none of the five algorithms erroneously rejected any VCV instances when provided the location of the manual boundary. When the RFF algorithm had to identify vocal cycles using the approximate midpoint of the voiceless consonant, the algorithm using YIN (RMSE = 0.39 ST) and RAPT (MBE = 0.02 ST) for f_o estimation resulted in the least error. However, implementing RAPT and YIN each resulted in the largest number of erroneous rejections (72 offset and 94 onset for RAPT; 59 offset and 100 onset for YIN). In considering RMSE, MBE, and erroneous rejections together when provided the midpoint of the voiceless consonant as a reference, Auditory-SWIPE′ resulted in the best performance (RMSE = 0.43 ST;

TABLE III. Comparison of fundamental frequency (f_o) estimation methods when provided with the manually determined time point corresponding to the vocal cycle closest to the voiceless consonant, and when provided only with the midpoint of the voiceless consonant.

	Manual boundary cycle as reference				Voiceless consonant as reference			
	Error (semitones, ST)		Erroneous rejections		Error (ST)		Erroneous rejections	
Method of f_o estimation	RMSE	MBE	Offset	Onset	RMSE	MBE	Offset	Onset
Autocorrelation	1.06	0.50	0	0	0.43	0.09	54	90
Halcyon	0.81	0.03	0	0	0.41	0.06	60	92
Auditory-SWIPE'	0.52	−0.20	0	0	0.43	0.04	52	98
RAPT	1.97	−1.13	0	0	0.50	0.02	72	94
YIN	0.80	−0.19	0	0	0.39	0.05	59	100

MBE = 0.04 ST; 52 erroneous offset rejections, 98 erroneous onset rejections).

Analyzing the performance of each f_o estimation method when provided with indices for the manually determined boundary cycle was conducted to simulate the downstream effects of the acoustic features accurately identifying the boundary between the voiced and voiceless segments. Because of its high-tier performance in this scenario, coupled with its superior performance when only provided with the midpoint of the voiceless consonant, Auditory-SWIPE′ was selected for f_o estimation in the optimized version of the algorithm V.2.

B. Development of category-specific thresholds

1. Automated sample rejection

A ROC curve was constructed from examining the discriminatory ability of pitch strength to distinguish between valid and invalid (i.e., manually rejected) RFF instances. A total of 3474 VCV instances were used to construct the ROC curve: 3271 offset and 2854 onset instances were valid, whereas 203 offset and 620 onset instances were invalid. The resulting area under the ROC curve was 0.73 (95% confidence interval = 0.71–0.75). Using the maximum PLR (100% specificity, 4% sensitivity), a pitch strength threshold of 0.05 was selected as rejection criterion, wherein speech samples with a pitch strength of 0.05 or lower would be rejected prior to RFF cycle analysis. Out of 3474 VCV instances, 19 VCV instances did not make this cutoff. Further analysis thus includes 3270 offset instances and 2853 onset instances that were not excluded due to manual rejection or low (<0.05) pitch strength values.

2. Boundary cycle shifts

Figure 4 shows the relationship between acoustic features and the true boundary cycle for the training dataset (3270 offset instances, 2853 onset instances). Mean acoustic feature values are shown as a function of the average number of pitch periods away from the true boundary cycle. The true boundary corresponds to the boundary cycle (i.e., offset cycle 10 for voice offset and onset cycle 1 for voice onset) that is selected by trained technicians during manual RFF analysis. Normalized peak-to-peak amplitude increased toward the voiced sonorant for both voice offset [Fig. 4(a); negative pitch period distance] and onset [Fig. 4(d); positive pitch period distance], yet was relatively stable during the voiceless consonant. Conversely, the number of zero crossings increased toward the voiceless consonant for voice offset [Fig. 4(b)] and voice onset [Fig. 4(e)], then was stable during the voiced sonorant. Waveform shape similarity—calculated in reference to the voiceless consonant—matched the trends observed in the number of zero crossings for voice offset [Fig. 4(c)] and voice onset [Fig. 4(f)]. Leveraging the relationship between acoustic feature values and the average number of pitch periods away from the true boundary cycle, effect sizes were calculated to identify feature-based boundary cycle shifts; resulting boundary cycle shifts are shown in Table VII in the Appendix.

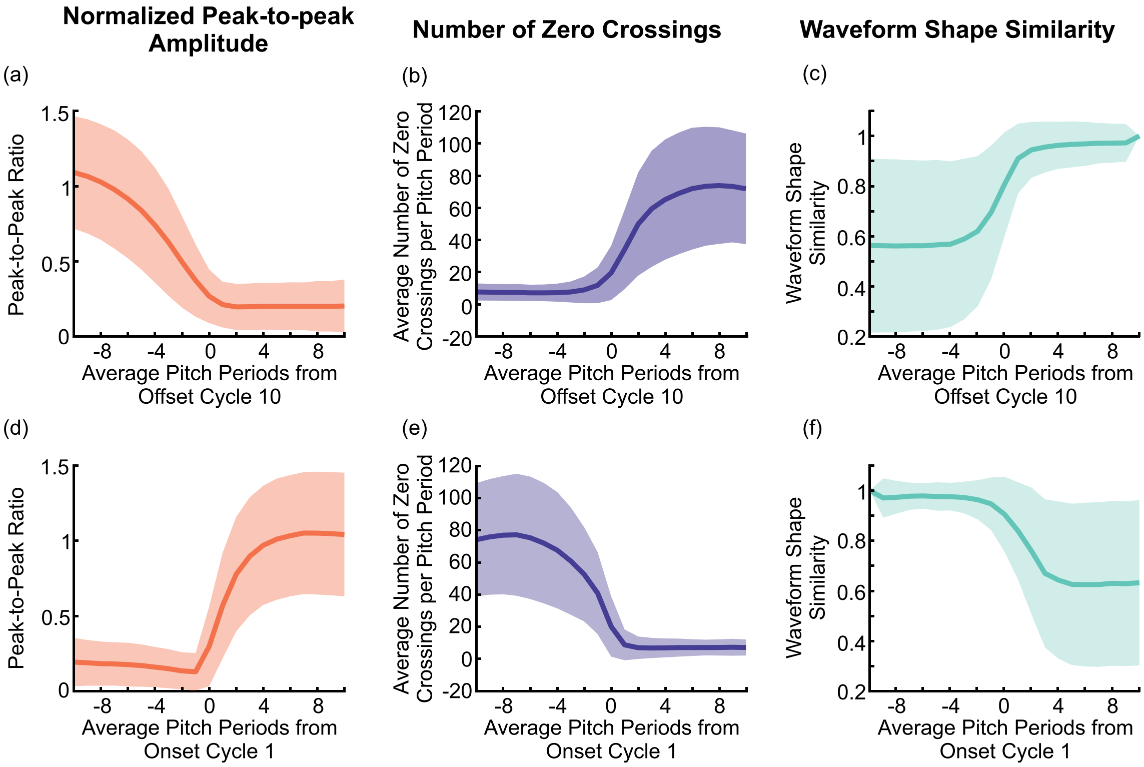

FIG. 4. (Color online) Mean acoustic feature values as a function of average pitch periods away from the true boundary. Voice offset is shown in (a)–(c), wherein the true boundary cycle is offset cycle 10. Voice onset is shown in (d)–(f), wherein the true boundary cycle is onset cycle 1. Trends for normalized peak-to-peak amplitude (orange) are shown in (a) and (d), number of zero crossings (purple) in (b) and (e), and waveform shape similarity (teal) in (c) and (f). Shading indicates 95% confidence intervals.

3. Category creation

Of the remaining 3270 offset instances and 2853 onset instances, 4 pitch strength cutoffs were selected via manual examination to describe the trends in acoustic feature values for voice offset: 0.15, 0.25, 0.35, and 0.45. Thus, in addition to the rejection criteria of 0.05, five pitch strength categories resulted for voice offset as follows:

$$\mathrm{cat}_{\mathrm{off}}(S) = \begin{cases} 1,\ 0.05 < S \leq 0.15 \\ 2,\ 0.15 < S \leq 0.25 \\ 3,\ 0.25 < S \leq 0.35 \\ 4,\ 0.35 < S \leq 0.45 \\ 5,\ S > 0.45. \end{cases} \tag{2}$$

In Eq. (2), pitch strength is denoted by the variable S, and the speech sample category is described by $\mathrm{cat}_{\mathrm{off}}$. Similar to voice offset, manual examination of the three features resulted in four pitch strength cutoffs for voice onset: 0.15, 0.25, 0.35, and 0.55. Five categories resulted for voice onset ($\mathrm{cat}_{\mathrm{on}}$) as a function of pitch strength (S).

Using these categories, optimal thresholds for each acoustic feature were determined (see Table VIII in the Appendix) and implemented into the V.3 algorithm. Figure 5 shows the performance

$$\mathrm{cat}_{\mathrm{on}}(S) = \begin{cases} 1,\ 0.05 < S \leq 0.15 \\ 2,\ 0.15 < S \leq 0.25 \\ 3,\ 0.25 < S \leq 0.35 \\ 4,\ 0.35 < S \leq 0.55 \\ 5,\ S > 0.55 \end{cases} \tag{3}$$

of V.1–V.3 algorithms in selecting the manually identified boundary cycle using speech samples of the training dataset. Out of 3270 instances to classify for voice offset [Figs. 5(a)–5(c)], the V.3 algorithm resulted in the largest number of correctly identified boundary cycles ($N = 1503$), followed by V.2 ($N = 1399$), and then V.1 ($N = 1349$). When considering the instances for which the automatically identified boundary cycle did not match the manually identified boundary cycle, the majority of misclassifications occurred closer to the voiced sonorant for V.1 ($N = 1692$), V.2 ($N = 1636$),

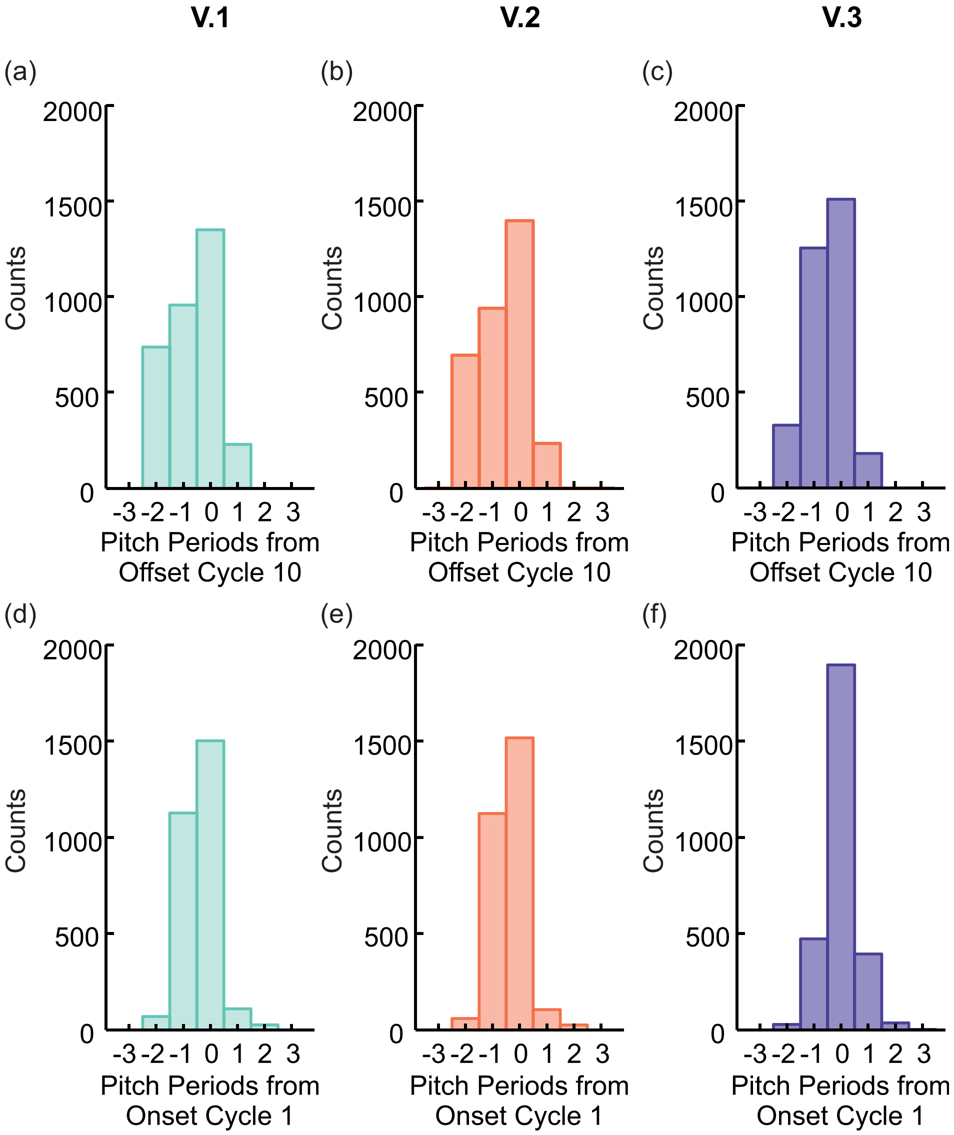

FIG. 5. (Color online) Boundary cycle identification by each of the semi-automated algorithms. Cycle classifications are measured as a function of average pitch periods from the manually identified boundary cycle. Voice offset is shown in (a)–(c), whereas voice onset is shown in (d)–(f). Results for V.1 (teal) are shown in (a) and (d), V.2 (orange) in (b) and (e), and V.3 (purple) in (c) and (f).

and V.3 ($N = 1584$). Out of 2853 instances to classify for voice onset [Figs. 5(d)–5(f)], V.3 resulted in the greatest number of correctly identified cycles ($N = 1896$). V.1 and V.2 produced similar results with 1502 correctly identified cycles for V.1 and 1517 for V.2. Dissimilar from voice offset, a great majority of misclassified boundary cycles were identified as occurring closer to the voiceless consonant for V.1 ($N = 1197$) and V.2 ($N = 1184$). However, results for V.3 showed a more even split for misclassified cycles: of the 937 misidentified cycles, the boundary cycle was identified as occurring closer to the voiced sonorant in 504 instances, whereas it was identified as being closer to the voiceless consonant in 434 instances.

As a next step, k-fold cross-validation was performed on all 3474 VCV instances to assess whether category and threshold parameters were overfit to the data. The cross-validation estimate of prediction error was averaged across $k = 10$ folds, resulting in an MBE of -0.03 ST (SD = 0.01 ST) and RMSE of 0.31 (SD = 0.004 ST) of the k-training set (N = 1042), and an MBE of -0.03 ST (SD = 0.04 ST) and RMSE of 0.32 ST (SD = 0.02 ST) in the k-validation set (N = 116). Given the small discrepancy between error estimates of the k-training and k-validation sets, it was determined that the constructed V.3 model was not overfit to the training data, and parameters were retained to finalize the algorithm.

C. RFF algorithm performance in the test set

Table IV shows the distribution of speech samples in the test set (873 VCV instances) by the pitch strength categories described in Eqs. (2) and (3); this distribution is described according to speaker sex, group (i.e., with versus without a voice disorder), and recording location (i.e., in a quiet room or waiting area versus a sound-attenuated room). When considering speaker sex, a larger proportion of female voices were rejected (1.9% for voice offset, 2.5% for voice onset) due to low pitch strength values than male voices (0% for voice offset, 0.7% for voice onset). When considering speaker group, a greater percentage of speech samples recorded from individuals with a voice disorder (2.6% for voice offset, 3.1% for voice onset) were rejected than samples recorded from individuals without a voice disorder (0% for voice offset, 0.9% for voice onset). Similarly, a greater proportion of samples recorded in a quiet room or waiting area (2.9% for voice offset, 3.3% for voice onset) were rejected compared to those recorded in a sound-attenuated room (0.4% for voice offset, 1.2% for voice onset), and more onset instances were rejected than offset instances for each factor. Of the 567 VCV instances corresponding to samples recorded in a sound-attenuated room, more than 50% of these instances (384 offset instances, 386 onset instances) were classified as having a pitch strength above 0.35 (i.e., categories 4 or 5). Of 306 VCV instances corresponding to samples recorded in a quiet room or waiting area of a voice clinic, the majority of these instances (179 offset instances, 152 onset instances) were classified as having a pitch strength below 0.35 (i.e., categories 1–3). To this end, a greater proportion of VCV instances from female speakers fell within the higher pitch strength categories (i.e., categories 4 and 5, with a pitch strength value greater than 0.35 for either voice offset or voice onset) than male speakers for both voice offset and onset (offset, female = 63.8%, male = 44.1%; onset, female = 64.8%, male = 52.0%). Similarly, a larger percentage of speakers without voice disorders (offset, 65.8%; onset, 66.0%) were characterized at these higher pitch strength categories than speakers without voice disorders (offset, 48.7%; onset, 55.1%) for offset and onset VCV instances. Finally, a greater proportion of speakers recorded in a sound-attenuated room resulted in higher pitch strength categories (offset, 67.7%; onset, 68.1%) than those recorded in a quiet room or waiting area (offset, 38.6%; onset, 47.1%).

RFF was computed for the independent test set (873 VCV instances) using each of the semi-automated algorithms. Table V shows the performance resulting from comparing each semi-automated RFF algorithm against manual RFF estimates in terms of MBE (ST) and RMSE (ST). When comparing MBE and RMSE between algorithm versions,

TABLE IV. Distribution of pitch strength categories for voice offset and voice onset instances in the test set (873 VCV instances from 291 speech samples). Values are shown as a percentage (%) of total VCV instances (N) and do not reflect speech samples that were rejected during pre-processing.

		Voice offset ($N = 286$)						Voice onset ($N = 290$)					
		Percent of samples per pitch strength category						Percent of samples per pitch strength category					
Speaker factor	N	0[a]	1	2	3	4	5	0[a]	1	2	3	4	5
Sex													
Male	279	0	2.9	12.5	40.5	35.1	9.0	0.7	9.3	11.8	26.2	49.8	2.2
Female	594	1.9	1.7	7.1	25.6	43.4	20.4	2.5	4.9	8.8	19.0	61.3	3.5
Group													
Voice disorder	423	2.6	3.3	12.1	33.3	35.5	13.2	3.1	7.3	11.1	23.4	54.1	0.9
No voice disorder	450	0	0.9	5.8	27.6	45.8	20.0	0.9	5.3	8.4	19.3	60.9	5.1
Location													
Quiet room	306	2.9	3.6	13.4	41.5	30.4	8.2	3.3	10.5	13.7	25.5	45.4	1.6
Sound booth	567	0.4	1.2	6.3	24.4	46.4	21.3	1.2	4.1	7.6	19.0	64.2	3.9

[a] A category of zero signifies that the instance was rejected because the pitch strength value was <0.05.

TABLE V. Comparison of manual and automated relative fundamental frequency estimates by algorithm version, computed using a test set of 291 speech samples. Error values are shown as mean (95% confidence interval).

Algorithm version	Mean bias error (ST)	RMSE (ST)
V.1	0.09 (0.07–0.11)	0.34 (0.32–0.36)
V.2	0.08 (0.05–0.10)	0.32 (0.29–0.34)
V.3	0.01 (-0.01–0.03)	0.28 (0.26–0.30)

V.3 results in the least error compared to manual RFF estimates (MBE = 0.01 ST, RMSE = 0.28 ST), followed by V.2 (MBE = 0.08 ST, RMSE = 0.32 ST), and then V.1 (MBE = 0.09 ST, RMSE = 0.34 ST). Figure 6 shows these errors by individual voice offset and onset cycles. On average, MBE of offset cycles 2–10 substantially decreases when using V.3 to calculate RFF compared to V.1 or V.2. The MBE of onset cycle 2 improves toward zero when using V.3, whereas that of onset cycle 1 and cycles 3–10 approach similar values across the three algorithm versions. Average RMSE also decreases for offset cycles 2–10 and onset cycle 1 when using V.3 with similar performance across V.1–V.3 algorithms for remaining cycles. Taking these findings into account, our results show that the V.3 semi-automated RFF estimation algorithm resulted in the greatest correspondence to manual RFF estimates.

Within the test set of 291 samples (873 VCV instances), MBE and RMSE values were examined across sample characteristics of signal acquisition quality and overall severity of dysphonia. The Welch's test examining MBE values across signal acquisition quality (i.e., recorded in a quiet room or waiting area versus sound-attenuated room) revealed that recording location produced a medium significant effect ($p = 0.04$, $d = 0.47$) on RFF values produced from the V.3 algorithm (Witte and Witte, 2010, p. 383). The average MBE was larger for sound samples recorded in a quiet room or waiting area ($M = 0.08$ ST, SD = 0.23 ST) compared to those recorded in a sound-attenuated room ($M = -0.02$ ST, SD = 0.21 ST). However, the Welch's test examining RMSE values across recording locations showed that recording location was not a significant factor ($p = 0.25$), with the average RMSE for sound samples recorded in a quiet room or waiting area ($M = 0.31$ ST, SD = 0.18 ST) similar to that of sound samples recorded in a sound-attenuated room ($M = 0.27$ ST, SD = 0.16 ST). Pearson product-moment correlation coefficients conducted for MBE and RMSE against overall severity of dysphonia elicited $r = -0.08$ ($p = 0.44$) and $r = 0.44$ ($p < 0.001$), respectively.

Because Auditory-SWIPE′ is a more computationally complex method than that of the autocorrelation method used in the V.1 algorithm, we compared the processing times necessary to compute the f_o contour of each test set speech sample when using the two f_o estimation techniques. On average, Auditory-SWIPE′ required 3.59 s (SD = 1.34 s) to process each speech sample containing three VCV instances, whereas autocorrelation required 0.28 s (SD = 0.11 s).

IV. DISCUSSION

In this study, we acquired a wide range of vocal signals to create a large RFF database that could be used to examine the impacts of the f_o estimation method and sample characteristics on resulting RFF estimates. We recorded a broad range of vocal function in a variety of locations, including clinic waiting areas, quiet rooms, and sound-attenuated rooms. Five f_o estimation methods were then evaluated to determine which method yielded the greatest correspondence to manual RFF estimates. The f_o estimation algorithm

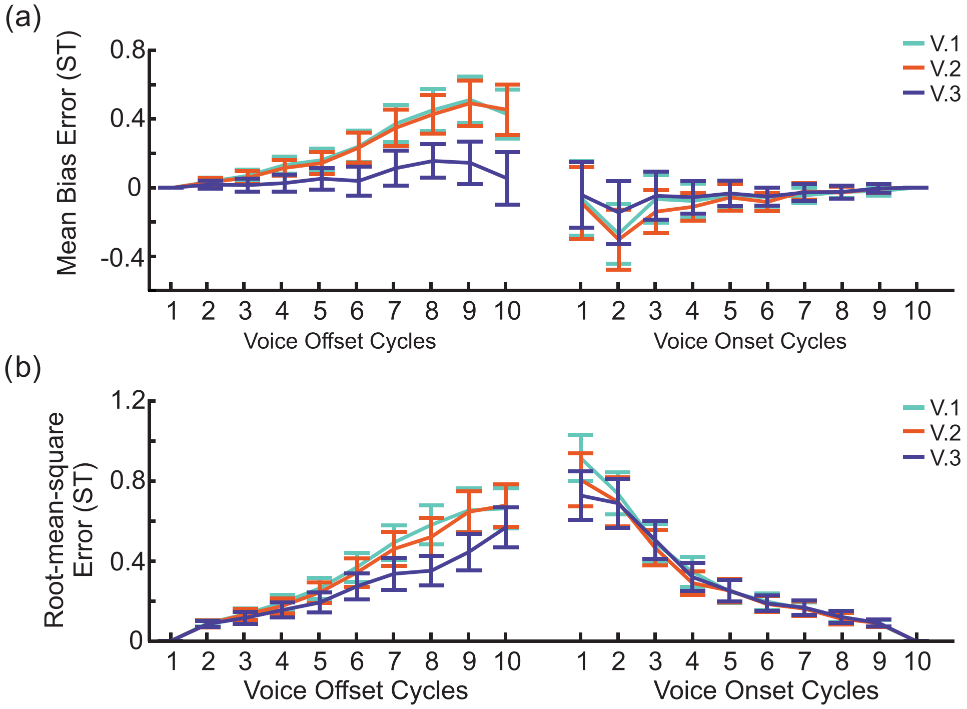

FIG. 6. (Color online) Mean bias error (MBE) (a) and RMSE (b) by voice offset and onset cycles for V.1 (teal), V.2 (orange), and V.3 (purple).

Auditory-SWIPE′ was implemented in the optimized, semi-automated RFF algorithm due to its superior performance when simulating the downstream effects of accurate boundary cycle identification. The effects of sample characteristics on RFF were then examined; we selected overall severity of dysphonia and signal acquisition quality to capture variation in sample characteristics (Lien *et al.*, 2017). We then used pitch strength as a means of objectively quantifying these characteristics. Using a training set from the RFF database, categories based on pitch strength values were developed, and RFF algorithm thresholds were tuned to each category. RFF values were then recalculated on a test set using the category-specific thresholds.

Our results show that refining the method of f_o estimation and accounting for sample characteristics leads to increased correspondence between manual and automated RFF estimates. Within this analysis, MBE and RMSE were both calculated, with MBE providing an approximation of average accuracy of the RFF estimates and RMSE providing insight into the precision of these estimates without regard to the direction of error (i.e., positive or negative ST). The MBE obtained after refining the algorithm was positive, suggesting that using the V.3 algorithm to compute RFF will, on average, generate a positively biased systematic error of approximately 0.01 ST. The average RMSE across samples in the test group was 0.28 ST; this indicates that the spread of error values will, on average, approach 0.28 ST when using the V.3 algorithm to estimate RFF. Despite the method of error computation, the refined version of the algorithm resulted in the least error when compared to the algorithm without modifications (i.e., using autocorrelation for f_o estimation) and when only optimizing f_o estimation method (i.e., using Auditory-SWIPE′ for f_o estimation). Thus, although the Auditory-SWIPE′ algorithm is more computationally complex than autocorrelation, wherein more processing time is required to compute f_o, the trade-off for more accurate f_o estimation is necessary in order to improve the accuracy of the algorithm. Yet, it is important to note that the Auditory-SWIPE′ algorithm not only computes the f_o contour but also the pitch contour used to categorize speech samples. Overall, our results suggest that identifying the voiced/unvoiced boundary using pitch strength-tuned thresholds results in a greater correspondence to manual estimates of RFF across a broad range of vocal function.

Rather than assessing RFF across specific sample characteristics, such as clinical diagnosis, we included a myriad of diagnoses (see Table VI in the Appendix) in order to increase correspondence between manual and semi-automated RFF algorithms across the spectrum of vocal function. We used pitch strength as a gestalt estimate of acoustic voice quality, thereby encompassing specific sample characteristics of overall severity of dysphonia and signal acquisition quality. However, it is important to consider how RFF estimates computed using V.3 vary across clinical metrics. By examining errors across recording location, we found that recording location was not a significant factor in the model for RMSE. This suggests that the average spread of RFF estimates was relatively similar for speech samples recorded in a quiet room/waiting area or sound-attenuated room. These findings indicate that the precision of RFF estimates is not affected by signal acquisition quality. Conversely, recording location was a significant factor in the model for MBE, wherein the average accuracy of RFF estimates was reduced for samples recorded in a quiet room or waiting area. These findings suggest that the bias of RFF values is still affected by signal acquisition quality despite using pitch strength to account for clinical sample characteristics. Our results are indicative of systematic errors to occur, on average, on the order of 0.08 ST for samples recorded in a quiet room or waiting area and −0.02 ST for samples recorded in a sound-attenuated room.

When examining RFF errors across overall severity of dysphonia, a very weak, negative relationship (Evans, 1996, p. 146) was found between overall severity of dysphonia and MBE. These results indicate that, dissimilar from findings for signal acquisition quality, the accuracy of an RFF estimate from the V.3 algorithm will not be substantially altered from manual estimates as a function of overall severity of dysphonia. On the other hand, a moderate, positive relationship (Evans, 1996, p. 146) was found between overall severity of dysphonia and RMSE; this suggests that the precision of resulting RFF values may be positively related to the overall severity of dysphonia of the speaker. Thus, although we used pitch strength to acoustically account for the clinical metrics of signal acquisition quality and dysphonia severity, pitch strength may not be sufficient to fully encompass these perceived sample characteristics.

In the current study, we optimized the semi-automated RFF algorithm to increase correspondence with manual RFF; however, neither the average MBE nor the RMSE between manual and V.3 RFF estimates were zero. Possible reasons for this outcome are twofold. First, as previously mentioned, pitch strength may not be sufficient to account for differences in signal acquisition quality and/or dysphonia severity present in current clinical practice. As such, future investigations should investigate additional or alternative acoustic metrics to account for the diversity in these clinical sample characteristics. Examples of such metrics may include cepstral peak prominence to assess speaker-related sample characteristics (Anand *et al.*, 2018), and/or signal-to-noise ratio to assess environmental-related sample characteristics. Second, it is unclear as to whether manual RFF is a true gold standard; therefore, it may not be necessary to remove errors between automated and manual RFF estimates. Manual RFF is derived using microphone signals (Eadie and Stepp, 2013; Goberman and Blomgren, 2008; Robb and Smith, 2002; Stepp, 2013; Stepp *et al.*, 2010; Stepp *et al.*, 2011b; Stepp *et al.*, 2012; Watson, 1998; Watson and Schlauch, 2008) or accelerometer signals (Lien *et al.*, 2015a). However, there may be a discrepancy between using these signals and the physiological initiation or termination of voicing at the vocal fold level. Trained technicians exercise trial-and-error to identify this physiological boundary via manual RFF estimation. Due to the subjective nature of this process, the selected boundary may not be the true initiation or termination of voicing. The semi-automated RFF algorithm makes use of three acoustic features—normalized peak-to-peak amplitude, number of zero crossings, and waveform shape similarity—to identify this transition point between voiced and unvoiced segments. Yet, it is unclear as to how these features relate to the physiological vibrations of the vocal folds during the transition into and out of voicing. As a result, investigation into the physiological relevance of RFF via manual and semi-automated techniques is warranted.

It is important to consider how the errors between manual and automated RFF values compare to meaningful differences in RFF values reported in the literature. For instance, after undergoing voice therapy, individuals with vocal hyperfunction produced increased RFF values, similar to those seen in healthy controls (Stepp *et al.*, 2011b; Stepp *et al.*, 2010). The largest changes were observed in the RFF cycles surrounding the voiceless consonant in a VCV utterance: average RFF values increased after voice therapy by +0.5 ST for voice offset cycle 10 and +0.81 ST for voice onset cycle 1. Notably, the average accuracy of RFF estimates when using the V.3 algorithm was found to be +0.05 ST for voice offset cycle 10 and -0.04 ST for voice onset cycle 1. These results suggest that the MBE associated with using the V.3 algorithm is on the order of one magnitude smaller than the increases in RFF that were observed in Stepp *et al.* (2011b). Thus, users can expect that, on average, clinically meaningful changes in RFF will not be masked by errors associated with using the V.3 algorithm to compute RFF (i.e., instead of manual estimation techniques).

Although the current study details preliminary steps taken to refine the semi-automated algorithm for RFF estimation, further investigation is warranted to continue to enhance accuracy and versatility across a broad range of vocal function. Specifically, the sample distribution analyzed in this study may not be fully representative of clinical practice. For instance, Martins *et al.* (2016) reports a substantial prevalence of vocal polyps in adults with voice disorders (12% of 2019 adults analyzed); however, only 3% of the population examined in the current study was diagnosed with vocal polyps (see Table VI in the Appendix). Furthermore, nearly 37% of the speakers with voice disorders analyzed in the current study were diagnosed with Parkinson's disease, and approximately 33% were diagnosed with muscle tension dysphonia. Because a substantial portion of our sample group consisted of these individuals, it is possible that our results are biased toward speakers with Parkinson's disease and speakers with muscle tension dysphonia. As such, future studies should take care to ensure that the prevalence of voice disorders in the examined population is representative of those seen in clinical practice. Doing so will enhance the clinical relevance of using RFF to acoustically examine vocal function. Additionally, it is unclear whether the heterogeneity of the equipment used to capture speech acoustics played a role in the differences seen in RFF accuracy in terms of signal acquisition quality. In particular, we hypothesized that signal acquisition quality was a feature of acoustic speech samples that affected the accuracy of RFF estimates; however, we examined signal acquisition quality solely in terms of whether the speech sample was recorded in a sound-attenuated room versus a quiet room or waiting area. As such, future work should also take into account the equipment used to record speech and the characteristics of the recording environment (e.g., background noise levels, reverberation) when examining signal acquisition quality.

V. CONCLUSIONS

RFF has shown promise as an acoustic measure for assessing and tracking vocal strain; however, semi-automated RFF is not yet transferable to the clinic due to instability across a wide range of vocal signals. Thus, we evaluated the impacts of f_o estimation method and sample characteristics on the correspondence between automated and gold-standard manual RFF estimates. Upon refining the f_o estimation method using the Auditory-SWIPE′ algorithm, in conjunction with accounting for sample characteristics via pitch strength categories, the accuracy and precision of semi-automated RFF estimates increased by 88.4% and 17.3%, respectively. These findings highlight the importance of considering the broad range of vocal function that may be encountered in clinical populations.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation under Grant No. 1247312 (J.M.V.) and the National Institutes of Health under Grant No. DC015570 (C.E.S.). The authors would like to thank Stephanie Lien, Carolyn Calabrese, and Carolyn Michener for assistance in data processing.

APPENDIX

See Tables VI–VIII for additional information regarding participant demographics (Table VI), acoustic features around the manually determined boundary cycle (Table VII), and resulting acoustic feature thresholds implemented in the refined RFF algorithm (Table VIII).

TABLE VI. Frequency of primary voice-related problems for speakers with disordered voices.

Primary voice-related problem	Frequency of problem
Acid reflux	3
Cyst(s)	3
Dysphagia	3
Ear, nose, and/or throat infection	3
Edema	4
Globus sensation	1
Granuloma	4
Laryngeal trauma	3
Muscle tension dysphonia	83
Nodules	20
Papilloma	2
Paradoxical vocal fold motion	1
Parkinson's disease	74
Polyp	6
Presbylarynges	1
Spasmodic dysphonia	6
Upper respiratory infection	1
Vocal fold atrophy	3
Vocal fold paralysis	2
Vocal fold scarring	4

TABLE VII. Automated boundary cycle computation as a function of average pitch periods from the manually determined (true) boundary cycle. Vocal cycles that elicit the maximum effect size are highlighted in gray, and were subsequently chosen as the automated boundary cycle.

	Voice offset			Voice onset		
Acoustic feature	d	Boundary cycle distance[a]	Boundary cycle direction[b]	d	Boundary cycle distance[a]	Boundary cycle direction[b]
Normalized peak-to-peak ratio	−1.06	1	−	0.63	1	+
	−0.87	0	−	1.29	0	+
	−0.54	1	−	1.44	1	+
Number of zero crossings	0.76	1	+	−1.20	1	−
	1.03	0	+	−1.36	0	−
	1.05	1	+	−0.76	1	−
Waveform shape similarity	0.66	1	+	−0.38	1	−
	0.78	0	+	−0.51	0	−
	0.64	1	+	−0.53	1	−

[a]Boundary cycle distance refers to the number of pitch periods away from the true boundary cycle.

[b]Boundary cycle direction corresponds to the direction (negative or positive) of the feature slope across the examined cycle; a positive shift indicates a shift forward in time, and a negative shift indicates a shift backward in time.

TABLE VIII. Optimal thresholds obtained at the Youden index from the ROC curves for voice offset and onset features. PTP = normalized peak-to-peak amplitude, NZC = number of zero crossings, and WSS = waveform shape similarity.

	Optimal thresholds by pitch strength (S) category									
	Voice offset					Voice onset				
Acoustic feature	$0.05 < S \leq 0.15$	$0.15 < S \leq 0.25$	$0.25 < S \leq 0.35$	$0.35 < S \leq 0.45$	$S > 0.45$	$0.05 < S \leq 0.15$	$0.15 < S \leq 0.25$	$0.25 < S \leq 0.35$	$0.35 < S \leq 0.55$	$S > 0.55$
PTP	0.1396	0.1002	0.1915	0.1773	0.1818	0.2244	0.1921	0.2310	0.1804	0.1277
NZC	18	24	17	11	10	20	13	13	11	5
WSS	0.8611	0.7842	0.8219	0.7648	0.6681	0.9640	0.9421	0.8840	0.9535	0.8361

[1] The V.1 algorithm is available for download at: http://sites.bu.edu/stepplab/research/rff/ (Last viewed May 30, 2019).

[2] Signal types were introduced as a classification scheme by Titze (1995) to recognize qualitative changes in voice signals. The three types of voice signals are described as follows: type 1 signals are nearly periodic signals, type 2 signals contain some bifurcations (e.g., alternating changes in fundamental frequency) such that there is no obvious single fundamental frequency throughout a segment, and type 3 signals have no apparent periodic structure.

[3] Gender information was not collected.

[4] We used overall severity of dysphonia to describe voice quality rather than the specific dimension of strain since strain is considered to be one of the least reliable and perceptually salient features of voice (Dejonckere *et al.*, 1996; Zraick *et al.*, 2011). We further support the use of overall severity of dysphonia—which provides a composite judgment of perceived dysphonia—as a speaker-related characteristic in this study because strain is thought to covary with other dimensions of voice quality such as roughness and breathiness (Zraick *et al.*, 2011).

[5] The dataset used to train individuals in manual relative fundamental frequency estimation is a separate dataset from that described here and may be downloaded from https://sites.bu.edu/stepplab/research/rff/ (Last viewed May 30, 2019).

Anand, S., Kopf, L. M., Shrivastav, R., and Eddins, D. A. (**2018**). "Objective indices of perceived vocal strain," J. Voice (published online).

Askenfelt, A. G., and Hammarberg, B. (**1986**). "Speech waveform perturbation analysis: A perceptual-acoustical comparison of seven measures," J. Speech Lang. Hear. Res. **29**(1), 50–64.

Awan, S. N., and Frenkel, M. L. (**1994**). "Improvements in estimating the harmonics-to-noise ratio of the voice," J. Voice **8**(3), 255–262.

Azarov, E., Vashkevich, M., and Petrovsky, A. (**2016**). "Instantaneous pitch estimation algorithm based on multirate sampling," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 20–25 March 2016, pp. 4970–4974.

Bainbridge, K. E., Roy, N., Losonczy, K. G., Hoffman, H. J., and Cohen, S. M. (**2017**). "Voice disorders and associated risk markers among young adults in the United States," Laryngoscope **127**(9), 2093–2099.

Baken, R. J., and Orlikoff, R. F. (**2000**). *Clinical Measurement of Speech and Voice* (Singular Thomson Learning, San Diego, CA).

Bhattacharyya, N. (**2014**). "The prevalence of voice problems among adults in the United States," Laryngoscope **124**(10), 2359–2362.

Bhuta, T., Patrick, L., and Garnett, J. D. (**2004**). "Perceptual evaluation of voice quality and its correlation with acoustic measurements," J. Voice **18**(3), 299–304.

Boersma, P. (**2001**). "Praat, a system for doing phonetics by computer," Glot Int. **5**(9-10), 341–345.

Camacho, A. (**2012**). "On the use of auditory models' elements to enhance a sawtooth waveform inspired pitch estimator on telephone-quality signals," in *2012 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA)*, pp. 1080–1085.

de Cheveigne, A., and Kawahara, H. (**2002**). "YIN, a fundamental frequency estimator for speech and music," J. Acoust. Soc. Am. **111**(4), 1917–1930.

Dejonckere, P. H., Bradley, P., Clemente, P., Cornut, G., Crevier-Buchman, L., Friedrich, G., Van De Heyning, P., Remacle, M., and Woisard, V. (**2001**). "A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques. Guideline elaborated by the Committee on Phoniatrics of the European Laryngological Society (ELS)," Eur. Arch. Otorhinolaryngol. **258**(2), 77–82.

Dejonckere, P. H., Remacle, M., Fresnel-Elbaz, E., Woisard, V., Crevier-Buchman, L., and Millet, B. (**1996**). "Differentiated perceptual evaluation of pathological voice quality: Reliability and correlations with acoustic measurements," Rev. Laryngol. Otol. Rhinol. (Bord). **117**(3), 219–224.

Deliyski, D. D., Shaw, H. S., and Evans, M. K. (**2005**). "Adverse effects of environmental noise on acoustic voice quality measurements," J. Voice **19**(1), 15–28.

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

Eadie, T. L., and Stepp, C. E. (**2013**). "Acoustic correlate of vocal effort in spasmodic dysphonia," Ann. Otol. Rhinol. Laryngol. **122**(3), 169–176.

Eddins, D. A., Anand, S., Camacho, A., and Shrivastav, R. (**2016**). "Modeling of breathy voice quality using pitch-strength estimates," J. Voice **30**(6), 774.e1–774.e7.

Evans, J. D. (**1996**). *Straightforward Statistics for the Behavioral Sciences* (Brooks/Cole, Pacific Grove, CA).

Gallena, S., Smith, P. J., Zeffiro, T., and Ludlow, C. L. (**2001**). "Effects of levodopa on laryngeal muscle activity for voice onset and offset in Parkinson disease," J. Speech Lang. Hear. Res. **44**(6), 1284–1299.

Goberman, A. E., and Blomgren, M. (**2008**). "Fundamental frequency change during offset and onset of voicing in individuals with Parkinson disease," J. Voice **22**(2), 178–191.

Hillman, R. E., Holmberg, E. B., Perkell, J. S., Walsh, M., and Vaughan, C. (**1989**). "Objective assessment of vocal hyperfunction: An experimental framework and initial results," J. Speech Lang. Hear. Res. **32**(2), 373–392.

Hirano, M. (**1981**). "Psycho-acoustic evaluation of voice," in *Clinical Examination of Voice*, edited by G. E. Arnold, F. Winckel, and B. D. Wyke (Springer, Wien), Vol. 5, pp. 81–84.

Jouvet, D., and Laprie, Y. (**2017**). "Performance analysis of several pitch detection algorithms on simulated and real noisy speech data," In *2017 25th European Signal Processing Conference (Eusipco)*, pp. 1614–1618.

Kempster, G. B., Gerratt, B. R., Verdolini Abbott, K., Barkmeier-Kraemer, J. M., and Hillman, R. E. (**2009**). "Consensus auditory-perceptual evaluation of voice: Development of a standardized clinical protocol," Am. J. Speech Lang. Pathol. **18**(2), 124–132.

Kopf, L. M., Jackson-Menaldi, C., Rubin, A. D., Skeffington, J., Hunter, E. J., Skowronski, M. D., and Shrivastav, R. (**2017**). "Pitch strength as an outcome measure for treatment of dysphonia," J. Voice **31**(6), 691–696.

Kuhn, M., and Johnson, K. (**2013**). *Applied Predictive Modeling* (Springer, New York).

Lien, Y. S. (**2015**). "Optimization and automation of relative fundamental frequency for objective assessment of vocal hyperfunction," Doctoral dissertation, Boston University, ProQuest Dissertations and Theses Global, available at https://open.bu.edu/bitstream/handle/2144/13645/Lien_bu_0017E_11638.pdf (1735392700) (Last viewed May 20, 2019).

Lien, Y. S., Calabrese, C. R., Michener, C. M., Heller Murray, E. S., Van Stan, J. H., Mehta, D. D., Hillman, R. E., Noordzij, J. P., and Stepp, C. E. (**2015a**). "Voice relative fundamental frequency via neck-skin acceleration in individuals with voice disorders," J. Speech Lang. Hear. Res. **58**(5), 1482–1487.

Lien, Y. S., Gattuccio, C. I., and Stepp, C. E. (**2014**). "Effects of phonetic context on relative fundamental frequency," J. Speech Lang. Hear. Res. **57**, 1259–1267.

Lien, Y. S., Heller Murray, E. S., Calabrese, C. R., Michener, C. M., Van Stan, J. H., Mehta, D. D., Hillman, R. E., Noordzij, J. P., and Stepp, C. E. (**2017**). "Validation of an algorithm for semi-automated estimation of voice relative fundamental frequency," Ann. Otol. Rhinol. Laryngol. **126**(10), 712—716.

Lien, Y. S., Michener, C. M., Eadie, T. L., and Stepp, C. E. (**2015b**). "Individual monitoring of vocal effort with relative fundamental frequency: Relationships with aerodynamics and listener perception," J. Speech Lang. Hear. Res. **58**(3), 566–575.

Lien, Y. S., and Stepp, C. E. (**2013**). "Automated estimation of relative fundamental frequency," in *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, Osaka, Japan, pp. 2136–2139.

Löfqvist, A., Baer, T., McGarr, N. S., and Story, R. S. (**1989**). "The cricothyroid muscle in voicing control," J. Acoust. Soc. Am. **85**(3), 1314–1321.

Lowell, S. Y., Kelley, R. T., Awan, S. N., Colton, R. H., and Chan, N. H. (**2012**). "Spectral- and cepstral-based acoustic features of dysphonic, strained voice quality," Ann. Otol. Rhinol. Laryngol. **121**(8), 539–548.

Ludlow, C. L. (**2009**). "Treatment for spasmodic dysphonia: Limitations of current approaches," Curr. Opin. Otolaryngol. Head Neck Surg. **17**(3), 160–165.

Martins, R. H., do Amaral, H. A., Tavares, E. L., Martins, M. G., Goncalves, T. M., and Dias, N. H. (**2016**). "Voice disorders: Etiology and diagnosis," J. Voice **30**(6), 761.e1–61.e9.

Maryn, Y., and Weenink, D. (**2015**). "Objective dysphonia measures in the program Praat: Smoothed cepstral peak prominence and acoustic voice quality index," J. Voice **29**(1), 35–43.

McKenna, V. S., and Stepp, C. E. (**2018**). "The relationship between acoustical and perceptual measures of vocal effort," J. Acoust. Soc. Am. **144**(3), 1643–1658.

Mehta, D. D., and Hillman, R. E. (**2008**). "Voice assessment: Updates on perceptual, acoustic, aerodynamic, and endoscopic imaging methods," Curr. Opin. Otolaryngol. Head Neck Surg. **16**(3), 211–215.

Morrison, M. D., Nichol, H., and Rammage, L. A. (**1986**). "Diagnostic criteria in functional dysphonia," Laryngoscope **96**(1), 1–8.

Poburka, B. J., Patel, R. R., and Bless, D. M. (**2017**). "Voice-vibratory assessment with laryngeal imaging (VALI) form: Reliability of rating stroboscopy and high-speed videoendoscopy," J. Voice **31**(4), 513e1–513e14.

Quatieri, T. F. (**2008**). *Discrete-Time Speech Signal Processing: Principles and Practice* (Prentice Hall, Upper Saddle River, NJ).

Rabiner, L. R. (**1977**). "Use of autocorrelation analysis for pitch detection," IEEE Trans. Acoust. Speech Signal Process. **25**(1), 24–33.

R Core Team (**2013**). R: A language and environment for statistical computing (R Foundation for Statistical Computing, Vienna, Austria), http://www.R-project.org/ (Last viewed October 7, 2019).

Reitermanova, Z. (**2010**). "Data splitting," in *WDS'10 Proceedings of Contributed Papers: Part I—Mathematics and Computer Sciences* (Matfyzpress, Prague), pp. 31–36.

Robb, M. P., and Smith, A. B. "Fundamental frequency onset and offset behavior: A comparative study of children and adults," J. Speech Lang. Hear. Res. **45**(3), 446–456 (**2002**).

Roy, N., Barkmeier-Kraemer, J., Eadie, T., Sivasankar, M. P., Mehta, D., Paul, D., and Hillman, R. (**2013**). "Evidence-based clinical voice assessment: A systematic review," Am. J. Speech Lang. Pathol. **22**(2), 212–226.

Roy, N., Fetrow, R. A., Merrill, R. M., and Dromey, C. (**2016**). "Exploring the clinical utility of relative fundamental frequency as an objective measure of vocal hyperfunction," J. Speech Lang. Hear. Res. **59**(5), 1002–1017.

Roy, N., Ford, C. N., and Bless, D. M. (**1996**). "Muscle tension dysphonia and spasmodic dysphonia: The role of manual laryngeal tension reduction in diagnosis and management," Ann. Otol. Rhinol. Laryngol. **105**(11), 851–856.

Roy, N., Merrill, R. M., Gray, S. D., and Smith, E. M. (**2005**). "Voice disorders in the general population: Prevalence, risk factors, and occupational impact," Laryngoscope **115**(11), 1988–1995.

Schwartz, S. R., Cohen, S. M., Dailey, S. H., Rosenfeld, R. M., Deutsch, E. S., Gillespie, M. B., Granieri, E., Hapner, E. R., Kimball, C. E., Krouse, H. J., McMurray, J. S., Medina, S., O'Brien, K., Ouellette, D. R., Messinger-Rapport, B. J., Stachler, R. J., Strode, S., Thompson, D. M., Stemple, J. C., Willging, J. P., Cowley, T., McCoy, S., Bernad, P. G., and Patel, M. M. (**2009**). "Clinical practice guideline: Hoarseness (dysphonia)," Otolaryngol. Head Neck Surg. **141**, 1–31.

Shrivastav, R., Eddins, D. A., and Anand, S. (**2012**). "Pitch strength of normal and dysphonic voices," J. Acoust. Soc. Am. **131**(3), 2261–2269.

Stepp, C. E. (**2013**). "Relative fundamental frequency during vocal onset and offset in older speakers with and without Parkinson's disease," J. Acoust. Soc. Am. **133**(3), 1637–1643.

Stepp, C. E., Heaton, J. T., Braden, M. N., Jette, M. E., Stadelman-Cohen, T. K., and Hillman, R. E. (**2011a**). "Comparison of neck tension palpation rating systems with surface electromyographic and acoustic measures in vocal hyperfunction," J. Voice **25**(1), 67–75.

Stepp, C. E., Hillman, R. E., and Heaton, J. T. (**2010**). "The impact of vocal hyperfunction on relative fundamental frequency during voicing offset and onset," J. Speech Lang. Hear. Res. **53**, 1220–1226.

Stepp, C. E., Merchant, G. R., Heaton, J. T., and Hillman, R. E. (**2011b**). "Effects of voice therapy on relative fundamental frequency during voicing offset and onset in patients with vocal hyperfunction," J. Speech Lang. Hear. Res. **54**(5), 1260–1266.

Stepp, C. E., Sawin, D. E., and Eadie, T. L. (**2012**). "The relationship between perception of vocal effort and relative fundamental frequency during voicing offset and onset," J. Speech Lang. Hear. Res. **55**(6), 1887–1896.

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

Talkin, D. (**1995**). "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, edited by W. B. Kleijn and K. K. Paliwal (Elsevier Science, New York), pp. 495–518.

Titze, I. R. (**1995**). "Workshop on acoustic voice analysis: Summary statement," in *National Center for Voice and Speech*, Denver, CO, available at http://www.ncvs.org/freebooks/WorkshopOnAcousticVoiceAnalysisProceedings_1995.pdf (Last viewed May 20, 2019).

Van Den Berg, J. (**1958**). "Myoelastic-aerodynamic theory of voice production," J. Speech Lang. Hear. Res. **1**(3), 227–244.

Vojtech, J. M., and Heller Murray, E. S. (**2019**). "Tutorial for manual relative fundamental frequency (RFF) estimation using Praat," available at https://sites.bu.edu/stepplab/research/rff/ (Last viewed May 20, 2019).

Watson, B. C. (**1998**). "Fundamental frequency during phonetically governed devoicing in normal young and aged speakers," J. Acoust. Soc. Am. **103**(6), 3642–3647.

Watson, P. J., and Schlauch, R. S. (**2008**). "The effect of fundamental frequency on the intelligibility of speech with flattened intonation contours," Am. J. Speech Lang. Pathol. **17**(4), 348–355.

Witte, R. S., and Witte, J. S. (**2010**). *Statistics* (Wiley, Hoboken, NJ).

Yiu, E. M., Lau, V. C., Ma, E. P., Chan, K. M., and Barrett, E. (**2014**). "Reliability of laryngostroboscopic evaluation on lesion size and glottal configuration: A revisit," Laryngoscope **124**(7), 1638–1644.

Youden, W. J. (**1950**). "Index for rating diagnostic tests," Cancer **3**(1), 32–35.

Zraick, R. I., Kempster, G. B., Connor, N. P., Thibeault, S., Klaben, B. K., Bursac, Z., Thrush, C. R., and Glaze, L. E. (**2011**). "Establishing validity of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V)," Am. J. Speech Lang. Pathol. **20**(1), 14–22.