Automated Relative Fundamental Frequency Algorithms for Use With Neck-Surface Accelerometer Signals

*,†Matti D. Groll, *,†Jennifer M. Vojtech, †Surbhi Hablani, ‡,§,||,¶Daryush D. Mehta, †,**Daniel P. Buckley, **J. Pieter Noordzij, and *,†,**Cara E. Stepp, *†‡§||¶***Boston, Massachusetts*

Summary: Objective. Relative fundamental frequency (RFF) has been suggested as a potential acoustic measure of vocal effort. However, current clinical standards for RFF measures require time-consuming manual markings. Previous semi-automated algorithms have been developed to calculate RFF from microphone signals. The current study aimed to develop fully automated algorithms to calculate RFF from neck-surface accelerometer signals for ecological momentary assessment and ambulatory monitoring of voice.

Methods. Training a set of 2646 /vowel-fricative-vowel/ utterances from 317 unique speakers, with and without voice disorders, was used to develop automated algorithms to calculate RFF values from neck-surface accelerometer signals. The algorithms first rejected utterances with poor vowel-to-noise ratios, then identified fricative locations, then used signal features to determine voicing boundary cycles, and finally calculated corresponding RFF values. These automated RFF values were compared to the clinical gold-standard of manual RFF calculated from simultaneously collected microphone signals in a novel test set of 639 utterances from 77 unique speakers.

Results. Automated accelerometer-based RFF values resulted in an average mean bias error (MBE) across all cycles of 0.027 ST, with an MBE of 0.152 ST and −0.252 ST in the offset and onset cycles closest to the fricative, respectively.

Conclusion. All MBE values were smaller than the expected changes in RFF values following successful voice therapy, suggesting that the current algorithms could be used for ecological momentary assessment and ambulatory monitoring via neck-surface accelerometer signals.

Key Words: Relative fundamental frequency−Accelerometer−Vocal hyperfunction.

INTRODUCTION

Relative fundamental frequency (RFF) is a family of acoustic measures that captures changes in fundamental frequency (f_o) during the transition into and out of a voiceless consonant (eg, in a vowel−voiceless consonant−vowel (VCV), production). Specifically, the instantaneous voice f_o, which describes the vibratory rate of the vocal folds,[1] is extracted from the 10 voicing cycles immediately preceding and following the voiceless consonant. The changes observed in f_o during these transition periods are hypothesized to be the result of the interplay between laryngeal muscle tension,[2,3] aerodynamics,[4] and vocal fold kinematics.[5] Laryngeal muscle tension, in particular, is thought to transiently elevate in order to assist in inhibiting voicing before, during, and immediately after the voiceless consonant.[2,3] This increases in laryngeal muscle tension are hypothesized to contribute to increases in offset and onset RFF values.

However, previous work has shown that RFF is reduced during voice offset and onset for individuals with voice disorders characterized by excessive laryngeal tension, including Parkinson's disease (PD),[6,7] laryngeal dystonia,[8] and vocal hyperfunction.[9] It has been postulated that the observed decreases in RFF are a result of higher baseline laryngeal muscle tension in these populations.[9,10] Higher baseline laryngeal tension has been hypothesized to cause a relative decrease in the transient elevation of laryngeal tension before, during, and after a voiceless consonant, therefore resulting in decreased RFF values when compared to individuals with healthy voices.[11] One study determined a relationship between RFF and vocal effort, wherein healthy individuals could modulate their vocal effort to achieve RFF values that were similar to those observed in individuals with excessive laryngeal muscle tension.[12] Further, the RFF values of cycles preceding the voiceless consonant were found to be correlated to listeners' auditory perception of vocal effort in individuals with typical voices who modulated their vocal effort[13] and in individuals with vocal hyperfunction.[11] Given that vocal strain is defined as the "perception of excessive vocal effort",[14] these studies indicate that RFF may have the potential to quantitatively

Accepted for publication June 4, 2020.

This work was supported by the National Institutes of Health Grant R01DC015570 (CES), P50 DC015446 (REM), and T32DC013017 (CAM) as well as the National Science Foundation Grant 1247312 (JMV). Thanks to Manuel Diaz Cadiz for help with MATLAB coding and algorithm implementation. Thanks to Megan Lee for help with speech sample classification. Thanks to Carolyn Calabrese, Katharine Kolin, Stephanie Lien, Austin Luong, Carolyn Michener, Roxanne Segina, and Monique Tardif for assistance with manual RFF estimation.

From the *Department of Biomedical Engineering, Boston University, Boston, 02215, Massachusetts; †Department of Speech, Language and Hearing Sciences, Boston University, Boston, 02215, Massachusetts; ‡Center for Laryngeal Surgery and Voice Rehabilitation and MGH Institute of Health Professions, Massachusetts General Hospital, Boston, 02114, Massachusetts; §Department of Surgery, Harvard Medical School, Boston, 02144, Massachusetts; ||Program in Rehabilitation Sciences, MGH Institute of Health Professions, Boston, 02129, Massachusetts; ¶Speech and Hearing Bioscience and Technology Program, Division of Medical Sciences, Harvard Medical School, Boston, 02144, Massachusetts; and the **Department of Otolaryngology − Head and Neck Surgery, Boston University School of Medicine, Boston, 02118, Massachusetts.

Address correspondence and reprint requests to Matti D. Groll, Department of Biomedical Engineering, Boston University, 677 Beacon St., Boston, MA, 02215 E-mail: mgroll@bu.edu

Journal of Voice, Vol. ■■, No. ■■, pp. ■■−■■
0892-1997

https://doi.org/10.1016/j.jvoice.2020.06.001

assess vocal strain across individuals, as well as to track changes in vocal strain within an individual over time.

Despite interest in implementing software to calculate RFF for clinical and research applications, manual estimation is currently the gold-standard for computing RFF. Trained technicians calculate RFF by visually inspecting each RFF instance and making a subjective decision about where the boundary between voiced and unvoiced speech occurs in a VCV utterance. Acoustic software, such as Praat,[15] is then used to compute the reciprocal of each vocal cycle duration of the 10 cycles prior to and following the voiceless consonant, corresponding to the instantaneous f_o of each cycle. In order to generate a reliable RFF estimate, the trained technician must repeat this process on at least six RFF speech sequences, totaling approximately 20 to 40 minutes per estimate.[8]

Due to the time-intensive nature and rigorous training required for manual estimation, a semiautomated method for RFF estimation using microphone signals was developed.[16] The semi-automated method of RFF estimation operates in five steps: 1) the voiceless consonant and vowels within a VCV utterance are identified, 2) the f_o range of the vowels are determined via autocorrelation, 3) positive and negative peaks in amplitude that potentially correspond to vocal cycles near the voiceless consonant are identified, 4) the boundary between voiced and voiceless segments is identified via acoustic feature selection, and 5) RFF values are calculated. Within this process, the first step requires the user to confirm if the locations of the voiceless consonant were correctly identified; if the user does not agree with the locations identified by the RFF algorithms, the user may then manually select the approximate midpoint of the voiceless consonant. Manual and automated RFF are each calculated via Equation 1, wherein the instantaneous f_o of the 10 cycles preceding and following the voiceless consonant are each normalized to the approximate steady-state f_o of the nearest vowel (f_o^{ref}). For voice offset, this approximate steady-state f_o is that of offset cycle 1, whereas for voice onset, it is that of onset cycle 10, as shown in Figure 1.

$$RFF^i\ (\mathrm{ST}) = 39.86 \times \log_{10}\left(\frac{f_o^i}{f_o^{\mathrm{ref}}}\right) \tag{1}$$

These original algorithms were further optimized to account for the broad range of signal qualities that can be expected for microphone signals.[17] The optimized algorithms improved f_o estimation by using the Auditory Sawtooth Waveform Inspired Pitch Estimator – Prime [18-20] and incorporated thresholds that were based on the signal quality of the specific speech sample. These changes resulted in smaller errors in the semiautomated RFF estimates when compared to the standard manual RFF estimates. This work was an important step towards the implementation of RFF for clinical voice assessment.

Although the majority of studies on RFF have employed signals acquired via microphone,[8-10,21] there has been a growing interest in using neck-surface vibrations generated during speech for ecological momentary assessment (EMA) and ambulatory voice monitoring (eg,[22-30]). These vibrations correspond to the underlying physiological mechanisms of voice production, and can be measured noninvasively via accelerometry. Unlike microphones, skin-surface accelerometers are less sensitive to background noise[31] and aid in preserving speaker confidentiality since the accelerometer signal cannot be used to construct intelligible speech.[32] Additionally, accelerometer signals captured from below the larynx are easier to analyze compared to microphone signals; this is because the resonances of the respiratory system are relatively time-invariant compared to the continuously varying resonances caused by the movement of articulators (ie, tongue, jaw, lips) during speech production.[33] Since many voice problems are related to daily vocal behavior, prescribed therapy may be improved if hyperfunctional vocal behaviors can be noninvasively monitored throughout a day of prolonged voice use.[24] Additionally, EMA and ambulatory monitoring allow researchers to easily collect a large number of speech samples from a single individual's daily voice use without the need for multiple visits to a clinic or laboratory.

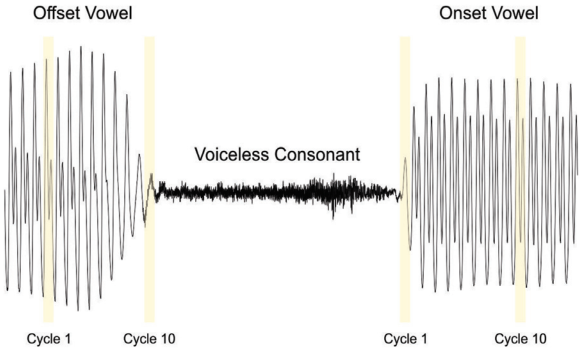

FIGURE 1. Microphone signal of a vowel–voiceless consonant –vowel utterance, /ifi/. The first and tenth vocal cycles of voice offset and voice onset have been identified. Instantaneous fundamental frequency (f_o) values of the 10 offset cycles and 10 onset cycles are normalized by offset cycle 1 and onset cycle 10, respectively.

It is important to consider how acoustic measures may differ when calculated from a microphone signal versus an accelerometer signal. Anterior neck-surface acceleration signals have been successfully used to derive voice characteristics related to RFF, average f_o, sound pressure level, vocal activity detection, phonation time, cepstral peak prominence, the relative amplitude of the first two harmonics (H1-H2), and glottal airflow features.[23,28,29,33-37] Previous studies have shown high correlations when comparing measures from microphone and accelerometer signals such as jitter and f_o in noisy environments.[23,35]

Although many acoustic measures are similar when derived from microphone and accelerometer signals, RFF measures are dependent on signal type. Studies have shown that manual RFF estimates from microphone and accelerometer signals both have the same general RFF pattern,[34,38] but that accelerometer-based RFF estimates were significantly lower for offset cycles.[34] Microphone signals are impacted by the radiation characteristics of the mouth and environmental noise, and accelerometer signals are impacted by neck surface transmission properties, which could result in inherent differences between the two signals.[28] Additionally, coarticulation of the voiceless consonant could cause masking noise in the microphone signal that would prevent identification of these cycles. Accelerometer-based vocal cycles were, on average, detected closer to the voiceless consonants when compared to microphone-based vocal cycles, and differences in resulting RFF estimates were reduced when vocal cycles were detected from a low-pass filtered version of the microphone signal to reduce the effects of coarticulation.[34] Thus, although manual RFF can be reliably estimated from accelerometer signals, they cannot be directly compared to manual RFF estimates from microphone signals.

Due to differences between microphone-based and accelerometer-based RFF estimates, the clinical relevance of accelerometer-based manual RFF estimates has not yet been shown to fully match the clinical relevance of microphone-based manual RFF estimates. One study showed that manual accelerometer-based RFF values of individuals with vocal hyperfunction were lower than those of individuals with healthy voices.[38] However, the clinical significance of manual microphone-based RFF estimates as, eg, a correlate for vocal effort,[13] a way to distinguish between different voice disorders,[6,39] and the ability to monitor changes following successful voice therapy,[10] has yet to be investigated in accelerometer-based manual RFF estimates. Therefore, microphone-based manual RFF estimates remain the gold-standard for clinical applications.

The limited clinical relevance for accelerometer-based manual RFF estimates is problematic because accelerometer signals have the potential to allow clinicians to assess vocal strain and track the progress of prescribed treatment via EMA or ambulatory voice monitoring. Additionally, accelerometer signals may provide a cleaner signal in typical clinical settings. Though microphone signals are almost universally accessible in clinics, accelerometers are less prone to environmental noise, which is a common problem in clinical settings.[38] Clinics have a noise level of 64.1 dBA,[40] exceeding recommendations that noise levels remain below 35 dBA when measuring acoustic signals with a headset microphone placed at a distance of 4 to 10 cm.[41] Thus, there is a need for a way to obtain accurate estimates of clinically relevant, gold-standard RFF measures (ie, those that would be obtained from microphone-based manual RFF calculations) from accelerometer signals alone. In the current study, we aimed to develop and evaluate automated RFF algorithms for neck-surface accelerometer signals. In order to investigate clinical applicability, the developed methodology was compared to gold-standard manual RFF estimates derived from microphone signals. Mean errors between the proposed algorithms and the gold-standard manual estimates were compared to the mean errors achieved by previous iterations of the RFF algorithms developed for microphone signals.[16,17]

METHODS

Participants and Recording Procedure

Participants were a selected subset of 394 speakers from the RFF database described in Vojtech et al, (2019). All participants selected had both microphone and neck-surface accelerometer data and had properly produced the required speech stimuli. All participants were fluent speakers of American English. Participants comprised 202 speakers without voice disorders (70 male, 132 female) aged 18 to 100 years (M = 35.4 years, SD = 21.4 years) and 192 speakers with voice disorders (61 male, 131 female) aged 18 to 84 years (M = 51.9 years, SD = 17.6 years). All speakers without voice disorders reported no prior history of speech, language, hearing or neurological disorders. Within the group of individuals with a voice disorder, individuals diagnosed with PD were diagnosed with idiopathic PD by a neurologist and were recorded while on their typical medication regimen. All other participants with a voice disorder (ie, muscle tension dysphonia, polyps, nodules, cysts, laryngeal dystonia) were diagnosed by a board-certified laryngologist. All participants completed written consent in compliance with the Boston University Institutional Review Board. A voice-specializing speech-language pathologist judged the overall severity of dysphonia (0-100) of each participant using the Consensus Auditory-Perceptual Evaluation of Voice.[14] The results of this assessment by voice group are shown in Table 1. In order to assess intrarater reliability for overall severity of dysphonia ratings, 15% of speech samples were reassessed by the same speech-language pathologist in a different sitting; Pearson's product-moment correlation coefficients were calculated via the statistical package R (Version 3.2.4), resulting in an intrarater reliability of $r = 0.96$.

Participants were recorded in one of the following environments: (1) in a waiting area or quiet room at Boston Medical Center using a dynamic headset microphone (model: WH20XLR; Shure, Niles, IL), (2) in a quiet room at Boston University using a condenser headset microphone (model: SM35XLR; Shure, Niles, IL), or (3) in a sound-attenuated room at Boston University using the same condenser headset microphone. For each environment, the headset microphone was placed 45° from the midline and 7 to 10 cm from the lips. Regardless of location, the accelerometer data was recorded by using an accelerometer sensor

TABLE 1.
Mean, Range, and Standard Deviation of the Perceptual Assessment of Overall Severity of Dysphonia in Speakers With and Without Voice Disorders

Voice group	Mean	Min	Max	Standard deviation
Without voice disorders	10.6	0	39.0	7.2
With voice disorders	22.4	0	100.0	20.7

(model BU-21771-000, Knowles, Illinois, USA). The sensor was placed on the surface of the neck just above the sternal notch and secured using medical grade adhesive (3M, St. Paul, MN). All microphone and accelerometer signals were sampled at 44.1 kHz with 16-bit resolution.

Each participant was instructed to produce a set of three uniform VCV utterances at their typical comfortable pitch and loudness. VCV utterances with the voiceless consonant, /f/, were selected for recording in order to minimize intraspeaker variability,[42] such that the selected VCV utterances were /afa/, /ifi/, and /ufu/. Participants were instructed to produce three /afa/ repetitions, take a breath, produce three /ifi/ repetitions, take a breath, and produce three /ufu/ repetitions. Each set of three repetitions was segmented into a separate speech sample. Speech samples with unusable utterances due to mispronunciation or errors in signal acquisition were removed from analysis. Each speaker produced an average of 8.3 usable VCV utterances across the three speech samples, resulting in 3285 VCV utterances from 1095 speech samples. Speech samples were separated into a training set and a test set, with 882 speech samples (2646 VCV utterances from 317 unique speakers) comprising the training set used to develop the algorithms and 213 speech samples (639 VCV utterances from 77 unique speakers) comprising the test set used to validate the algorithms.

Manual RFF Calculations

Manual RFF estimation was conducted for all VCV utterances by a minimum of two trained technicians (trained with an interrater reliability >0.93[1]) using Praat software. Manual analysis was performed using microphone signals, as this is the current gold-standard technique for RFF estimation. In order to manually calculate RFF within Praat, the f_o range was initially set to 90 to 500 Hz for female recordings and 60 to 300 Hz for male recordings; however, these settings were adjusted on an individual basis by the trained technician. The 10 voicing cycles on either side of the voiceless consonant /f/ were identified. The instantaneous f_o was computed as the inverse of the period of each voicing cycle. RFF was then calculated via Equation 1 for voice offset using the vocal cycles prior to the voiceless consonant and for voice onset using the vocal cycles following the voiceless consonant. For offset voicing cycles, RFF was computed relative to the instantaneous f_o of the first vocal cycle for voice offset (offset 1), whereas for onset voicing cycles, RFF was computed relative to the instantaneous f_o of the tenth vocal cycle (onset 10). An RFF instance was rejected if the trained technician determined that the sample was glottalized or misarticulated. Of the valid RFF instances, manual RFF values were averaged across repetition, set of utterances, and technician such that there was a single set of 10 RFF values for voice offset and 10 RFF values for onset per speaker. This averaged set was considered the gold-standard for that speaker when comparing the results of automated RFF algorithms.

TABLE 2.
Number of Speakers for Which Eight Trained Technicians Manually Computed Relative Fundamental Frequency. The Matrix Shows Common Speakers Analyzed Between Technicians, hereas the Diagonal (bolded) Describes the Number of Speakers a Single Technician Rated in Total

Technician	1	2	3	4	5	6	7	8
1	**223**							
2	63	**136**						
3	80	8	**91**					
4	67	0	0	**75**				
5	0	65	3	8	**77**			
6	0	0	0	0	0	**86**		
7	0	0	0	0	1	86	**87**	
8	13	0	0	0	0	41	41	**54**

Each technician reestimated 13% to 15% of their samples in a different sitting and the associated intrarater reliability was computed using Pearson's product-moment correlation coefficients. The average intrarater reliability was r = 0.91 (SD = 0.04, range = 0.87-0.99). Of the 394 speakers, 353 speakers were rated by two trained technicians and 41 speakers were rated by three trained technicians; Table II shows the breakdown of instances rated by each of eight trained technicians. Interrater reliability was calculated using an intraclass correlation coefficient (ICC), with an average interrater reliability of ICC (2,1) = 0.92 (SD = 0.06, range = 0.80-0.99).

Automated RFF Estimation

Previous semiautomated microphone-based RFF algorithms[2] were modified for automated accelerometer-based RFF estimation. Though the main analysis steps remain the same, several significant changes were made from previous algorithms based on the assumptions that ambulatory accelerometer samples are less affected by noise in the recording environment (ie, samples are not categorized by acoustic features such as pitch strength) and that ambulatory accelerometer samples can be collected in much larger quantities (ie, stricter rejection criteria may be implemented to ensure more reliable RFF estimates). The application of this methodology and the corresponding changes for use with neck-surface accelerometer signals are described here.

[1] 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/.

[2] The semi-automated RFF estimation algorithms for microphone signals can be downloaded from http://sites.bu.edu/stepplab/research/rff/.

Accelerometer Signal Quality Assessment

Accelerometer signals varied in quality based on sensor placement, skin properties and noise artifacts. In order to determine the quality of each accelerometer signal, the 882 speech samples in the training set were separated into two categories based on whether three VCV utterances could be distinguished from background noise in the accelerometer signals. Investigators visually inspected and listened to the accelerometer signals in order to identify VCV utterances. Speech samples in which three utterances could be identified in the accelerometer signal were categorized as “clear quality”, whereas those in which they could not were categorized as “poor quality.” Of the 882 speech samples, 690 were classified as clear quality, and 192 were classified as poor quality. This categorization was used to select algorithm rejection criteria as detailed in the following section.

Fricative Identification

Unlike previous semi-automated RFF estimates, the current fully automated accelerometer-based algorithms employed a method to detect the location of fricatives by assuming that each speech sample was comprised of three VCV utterances. This process is detailed in the Appendix. In short, the raw accelerometer signal was first band-pass filtered with a fifth-order Butterworth filter between 100 and 1000 Hz in order to remove low and high frequency noise in the accelerometer signal. As a result, parts of the signal with clear voicing were emphasized. The root mean square of this filtered signal was calculated over 300-point overlapping windows (10-point intervals). This RMS signal was then discretized to represent voiced and unvoiced regions of the signal. Additional steps (see Appendix) were used to remove regions of the signal that were incorrectly identified as voiced regions. The six longest-duration voiced regions were identified as the vowels surrounding each fricative. Fricative locations were then identified as the median point between each pair of these six vowel regions.

In order to ensure that the algorithms successfully identified vowels in the accelerometer signal, a ratio between the average RMS of the vowel sections ($\mathrm{RMS}_{\mathrm{OFFSET}}$ for offset and $\mathrm{RMS}_{\mathrm{ONSET}}$ for onset) and the average RMS of the silence sections ($\mathrm{RMS}_{\mathrm{SILENCE}}$), calculated from the two sections of the signal between the first and second VCV and the second and third VCV, was calculated based on Equation 2. If this ratio were below a set threshold, the accelerometer signal was determined to be too noisy to properly identify vowels, and the speech sample comprised of three VCV utterances was rejected from further analysis. This threshold was determined by a receiver operating characteristic (ROC) analysis to distinguish between the vowel-to-noise ratios of speech samples in the training set categorized as “clear quality” and the vowel-to-noise ratios of speech samples in the training set categorized as “poor quality.” A threshold of 5.6 was used in order to conservatively avoid “poor quality” samples at the expense of potentially rejecting some “clear quality” samples.

$$\text{Vowel-to-Noise Ratio} = \frac{1}{2}\left(\left(\frac{RMS_{OFFSET} - RMS_{SILENCE}}{RMS_{SILENCE}}\right) + \left(\frac{RMS_{ONSET} - RMS_{SILENCE}}{RMS_{SILENCE}}\right)\right) \quad (2)$$

f_o Estimation

Following the identification of the fricative locations, voicing cycle durations were calculated by estimating the f_o contour surrounding the fricatives. Three f_o estimation methods were tested to determine the optimal method for calculating changes in f_o on a cycle-by-cycle basis (autocorrelation, Auditory Sawtooth Waveform Inspired Pitch Estimator − Prime,[18-20] and Halcyon[43]). Each f_o estimation method was implemented into the current algorithms and used to evaluate the training set. The locations of the resulting automated RFF estimates were compared to the locations of the manual RFF estimates. Halcyon[43] led to the best correspondence between manual and automated RFF estimates. As a result, Halcyon was used in the final algorithms.

The f_o contours were used to identify potential cycle locations surrounding each fricative. Within each of these cycles, peaks were identified to refine the exact location of each cycle. Both positive and negative peaks were identified, and the set of peaks that was closer in time to the voiceless consonant was used for final RFF computation.

Boundary Cycle Identification

The output of the Halcyon f_o estimation was a vector of potential cycle locations. These cycle locations were used to determine which cycle corresponded to the boundary cycle where there was a voicing transition (ie, the final cycle of voicing in offset or the first cycle of voicing in onset). Unlike previous RFF algorithms, the current algorithms utilized different methods to identify the boundary cycle in the offset and onset vowels.

Acoustic features expected to change during voicing transitions (normalized peak-to-peak amplitude, number of zero crossings, and waveform shape-similarity) were used to identify the boundary cycle corresponding to the last cycle of voicing in offset vowels and the first cycle of voicing in onset vowels. In order to investigate how these signal features characteristically changed during voicing transitions, the cycle locations of manual RFF calculated from microphone signals were used to identify the gold-standard boundary cycle for each VCV utterance in the training set. The 10 voicing cycles preceeding the true offset boundary cycle and following the true onset boundary cycle were also identified using manual RFF. The average instantaneous f_o of each cycle was used to calculate the average cycle period and identify 10 additional potential cycle locations in the fricative, resulting in 10 potential cycle locations on either side of the true offset and onset boundary cycle. Signal features were calculated for each cycle and the resulting trends across the dataset were used to develop a method for

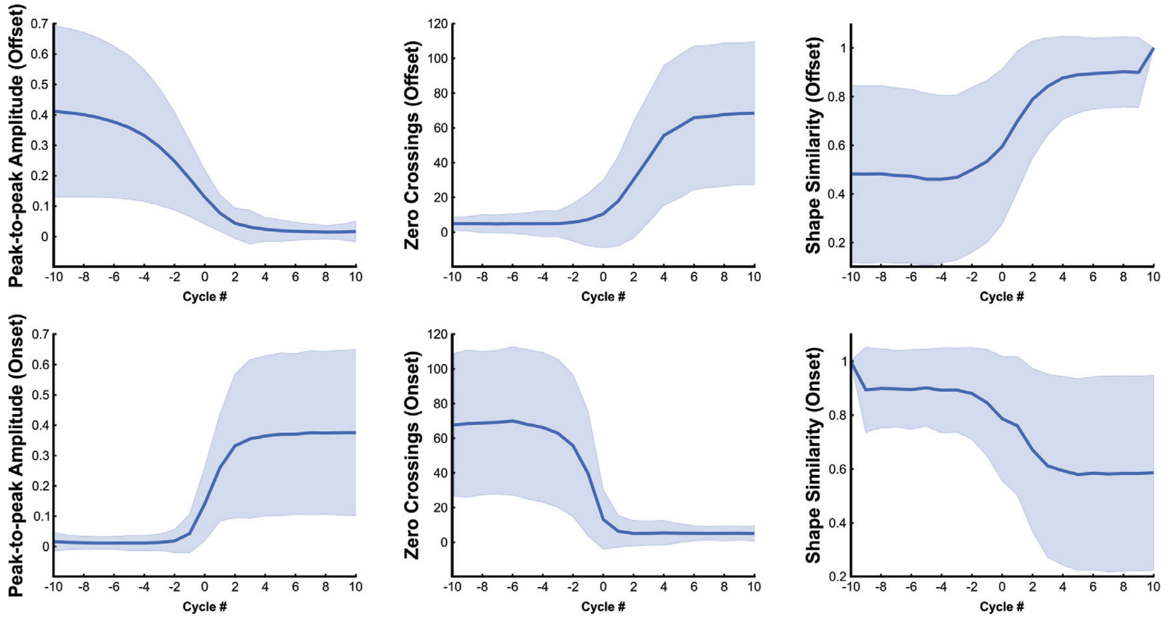

FIGURE 2. Average trends of acoustic features for each cycle surrounding the true voicing boundary (Cycle #0). Shaded regions indicate +/- 1 standard deviation from the average value at each cycle across the entire training set.

reliably distinguishing the true boundary cycle from the 10 surrounding cycles on either side.

The average trends of each acoustic feature for both offset and onset cycles are displayed in Figure 2, in which the 0th cycle corresponds to the true boundary cycle as identified by manual RFF. Positive cycles correspond to cycles following the true boundary cycle and negative cycles correspond to cycles preceding the true boundary cycle (eg, in offset, positive cycles are cycles located in the fricative and negative cycles are located in the vowel). Shaded regions indicate +/- 1 standard deviation. As expected, each feature behaved similarly in the accelerometer signal as in the microphone signal in previous algorithm development.[17] Specifically, normalized peak-to-peak values decreased during the fricative, whereas the number of zero crossings and shape similarity increased. Based on the large standard deviation of the shape similarity at all cycles, shape similarity was determined to be an unreliable feature for identifying the boundary location in accelerometer signals. The final implementation of this feature analysis for both offset and onset is summarized below. Further details are discussed in the Appendix.

Offset Boundary Cycle Identification

Methodology for identifying the boundary cycle that separated voiced and unvoiced segments during voicing offset was modified from techniques implemented in Lien et al (2017).[16] A sliding window traversed the accelerometer signal backward in time, from the midpoint of the fricative toward the vowel. The size of this window was determined by the average f_o of the VCV instance, calculated using the Halcyon f_o estimation algorithm. Within each window, peaks, and troughs in signal amplitude were collected. This sliding window process resulted in a vector of potential boundary cycle locations.

The following steps are summarized in Figure 3. Each potential boundary cycle location was considered independently as a single split point, such that the accelerometer signal was separated into a time-series vector representing a

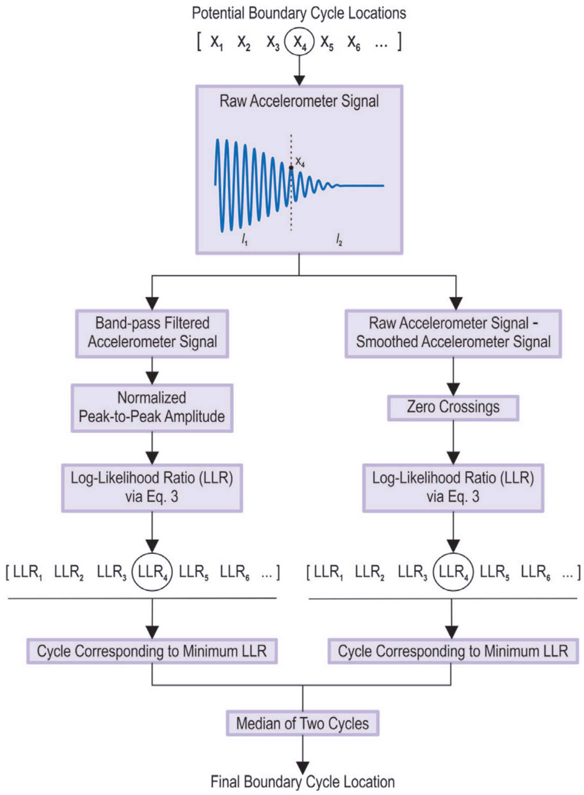
Potential Boundary Cycle Locations
[X1 X2 X3 X4 X5 X6 ...]
Raw Accelerometer Signal
X4
l1
l2
Band-pass Filtered Accelerometer Signal
Raw Accelerometer Signal - Smoothed Accelerometer Signal
Normalized Peak-to-Peak Amplitude
Zero Crossings
Log-Likelihood Ratio (LLR) via Eq. 3
Log-Likelihood Ratio (LLR) via Eq. 3
[LLR1 LLR2 LLR3 LLR4 LLR5 LLR6 ...]
[LLR1 LLR2 LLR3 LLR4 LLR5 LLR6 ...]
Cycle Corresponding to Minimum LLR
Cycle Corresponding to Minimum LLR
Median of Two Cycles
Final Boundary Cycle Location

FIGURE 3. Flowchart of steps used to calculate the offset boundary cycle location.

potential voiced segment to the left of the split point and another representing a potential unvoiced segment to the right of the split point. Normalized peak-to-peak amplitude and the number of zero crossings were calculated for the two segments. Prior to calculating the two acoustic features, the accelerometer signal was first preprocessed. Normalized peak-to-peak amplitude was calculated from a version of the accelerometer signal that was band-pass filtered using a second-order elliptic filter with cutoff frequencies equal to three semitones below the minimum and above the maximum f_o of the f_o contour. The number of zero crossings was calculated using a 1000-point smoothed version of the accelerometer signal that was subtracted from the raw accelerometer signal to remove low-frequency drift from the signal.

To identify how well the potential boundary cycle correctly separated the accelerometer signal into a vowel segment and voiceless fricative segment, a log-likelihood ratio was calculated using each acoustic feature of the two time-series vectors via Equation 3, where l_1 corresponds to the length of the first time-series vector and l_2 corresponds to the length of the second time-series vector. This resulted in a single log-likelihood ratio based on normalized peak-to-peak amplitude and a single log-likelihood ratio based on the number of zero crossings for that specific split point (ie, the potential boundary cycle location). This was repeated for every potential boundary cycle location such that each potential boundary cycle was used as a split point, resulting in two series of log likelihood ratios (one based on normalized peak-to-peak amplitude and another based on the number of zero crossings), with each log likelihood ratio in the series corresponding to a particular potential boundary cycle location that separated the signal into two segments. The median of the potential boundary cycle that resulted in the minimum log likelihood ratio for

$$\text{Log-Likelihood Ratio} = l_1\left(\ln\frac{Feature_{l1}}{l_1}\right) + l_2\left(\ln\frac{Feature_{l2}}{l_2}\right) \qquad (3)$$

normalized peak-to-peak amplitude and the potential boundary cycle that resulted in the minimum log likelihood ratio for the number of zero crossings was marked as the final calculated boundary cycle.

Onset Boundary Cycle Identification

Initially, the same method used to identify the voicing boundary in offset was employed in onset. However, in the training set, this method alone resulted in a boundary cycle that was, on average, farther into the vowel than the manual RFF estimate, resulting in an under-estimation of onset RFF values. Thus, an additional method was implemented to further shift the boundary cycle towards the true voicing boundary.

Exploratory analysis of the accelerometer data showed that there was a more dramatic cycle-by-cycle change in the normalized peak-to-peak amplitude when initiating voicing in onset than when terminating voicing in offset (Figure 2). As a result, onset voicing could be characterized by very little variation in normalized peak-to-peak amplitudes for potential cycles that are actually located within the fricative, followed by large changes in normalized peak-to-peak amplitudes for potential cycles immediately upon the initiation of voicing. In order to utilize this feature, the five potential cycles furthest into the fricative were used to calculate the mean and standard deviation of the normalized peak-to-peak amplitudes that corresponded to the known fricative location. Using these values, the normalized peak-to-peak amplitude of each cycle was converted into a z-score. The first cycle resulting in a z-score that exceeded a set threshold was determined as the voicing boundary location. This threshold was determined by systematically tuning the threshold to determine which value resulted in the largest number of optimal boundary locations. It was found that this method was best when considering only normalized peak-to-peak amplitude instead of also incorporating zero crossing and shape similarity. However, further conditional thresholds are used to refine the boundary location based on the number of zero crossings in abnormal samples. These thresholds and all other thresholds are discussed and reported in the Appendix.

Upon comparison of the voicing boundary calculated by the z-score method with the voicing boundary determined by manual RFF estimation, we found that in the training set, the z-score method tended to identify the onset boundary cycle before the manual RFF estimate. As a result, the final voicing boundary cycle was calculated as the median value of the boundary cycle determined by the log likelihood method and the boundary cycle determined by the z-score method.

Final RFF Calculation

In addition to the cycles that corresponded to the voicing boundaries, the nine cycles prior to the boundary cycle in offset (offset 10) and following the boundary cycle in onset (onset 1) were used to determine RFF values according to Equation 1. A final check was used to remove RFF values that were considered physiologically invalid based on criteria used to simulate RFF removal during manual estimation from glottalization and misarticulation. This final check resulted in the rejection of an offset or onset instance within an individual VCV utterance. These rejection criteria were identical to those from previous semi-automated RFF algorithms and can be found in the code for previous iterations of the algorithms made available online.[16,17] Thus, final RFF values were reported for all utterances that cleared two stages of auto-rejection: the overall signal quality assessment (see *Fricative Identification*) and the final RFF check.

Model Performance

Automated/semiautomated RFF estimates were computed for an independent test set (213 speech samples from 77 speakers, totaling 639 VCV utterances prior to auto-rejection) using each of the following algorithms: (a) semi-automated RFF estimates computed using autocorrelation for f_o estimation in microphone signals[16] (henceforth referred to as "MIC-original algorithms"), (b) refined semi-automated RFF estimates using microphone signals[17] ("MIC-refined algorithms"), and (c) the current automated RFF estimates using accelerometer signals ("ACC-refined algorithms"). The mean bias error (MBE) and root mean square error (RMSE) between automated and manual RFF estimates were compared among the three algorithms. Additionally, Pearson's correlation coefficients (r) between automated and manual RFF estimates were compared among the three algorithms in order to capture orthogonal errors to MBE and RMSE.

Post hoc exploratory analyses were used to explore specific instances in which there was a large RMSE value between the automated ACC-refined algorithms estimates and the manual RFF estimates. Instances in which the RMSE was larger than 1.0 ST for offset 10 or onset 1 were visually inspected to determine whether the automated algorithms correctly identified the boundary cycle based on the manual RFF estimates. Boundary cycles that were incorrectly identified were determined to be either a result of signal quality or a pure algorithmic error by inspecting both the microphone and accelerometer signal. If the boundary cycle was correctly identified, then the error was attributed to differences in the f_o contours of the microphone and accelerometer signals.

RESULTS

Test Set Performance

A total of 312 VCV utterances (48.8%) were rejected in the ACC-refined algorithms due to signal quality assessment. An additional 88 offset and 118 onset utterances were removed in the final RFF check due to physiologically invalid RFF patterns. In comparison, 223 offset and 240 onset utterances were removed due to physiologically invalid RFF patterns in the MIC-original algorithms, and 340 offset and 311 onset utterances were removed due to physiologically invalid RFF patterns in the MIC-refined algorithms. In total, ACC-refined algorithms rejection criteria resulted in 30 participants with zero usable offset utterances and 33 participants with zero usable onset utterances, compared to the MIC-original algorithms that had three and four participants and the MIC-refined algorithms that had 10 and five participants with zero usable offset and onset utterances, respectively.

MBE and RMSE values were calculated by comparing manual RFF estimates to each of the three semi-automated/automated RFF values and are shown in Table 3. The MBE and RMSE values for offset 10 and onset 1 are specifically shown, because these cycles are most likely to be affected by changes in vocal function.[8,10,12,39,44]

Similarly, Pearson's correlation coefficients were calculated across all cycles, as well as for offset 10 and onset 1. ACC-refined algorithms resulted in an average Pearson's r value of 0.90, and Pearson's r values of 0.86 and 0.46 for offset 10 and onset 1, respectively. In comparison, MIC-original algorithms had an average Pearson's r value of 0.88, and Pearson's r values of 0.84 and 0.53 for offset 10 and onset 1, respectively. MIC-refined algorithms had an average Pearson's r value of 0.88, and Pearson's r values of 0.76 and 0.61 for offset 10 and onset 1, respectively.

Post Hoc Error Analysis

A post hoc analysis of the test set was used to identify reasons why individual RFF instances resulted in large RMSE errors. All utterances in which the RMSE was greater than 1.0 ST were visually inspected. Of the 88 offset and 111 onset utterances with RMSE greater than 1.0 ST, 53.4% of offset utterances and 43.2% of onset utterances were from speakers with voice disorders, comparable to 49.3% of all utterances in the test set that were from speakers with voice disorders. The locations of manual RFF cycles and automated RFF cycles were compared in both offset and onset. Each utterance was classified into one of four categories based on why the utterance resulted in such a large difference between the manual and automated RFF. These categories are described below. Sample instances of each category are shown in Figure 4.

Category 1: Utterances in which the microphone signal was masked by high-energy noise caused by coarticulation of the fricative and vowel resulted in automatic RFF estimates identifying voicing cycles further into the fricative than the manual RFF estimates. In these instances, lowpass filtering the microphone revealed additional voicing cycles that could not be identified during manual RFF of

TABLE 3.
Performance Metrics for Each of the Three Algorithms on A Test Set of 213 Speech Samples From 77 Speakers, Totaling 639 Vowel-Consonant-Vowel Utterances. Mean Bias Errors (MBE) and Root Mean Square Errors (RMSE) are Shown as Averages Across All 10 Voicing Cycles, as Well as Individually for Offset Cycle 10 and Onset Cycle 1

Algorithms Version	Average MBE (ST)	Average RMSE (ST)	Offset 10 MBE (ST)	Onset 1 MBE (ST)	Offset 10 RMSE (ST)	Onset 1 RMSE (ST)
MIC-original	0.11	0.30	0.53	0.08	0.71	0.86
MIC-refined	0.09	0.27	0.04	0.02	0.56	0.86
ACC-refined	0.03	0.30	0.15	-0.25	0.65	1.07

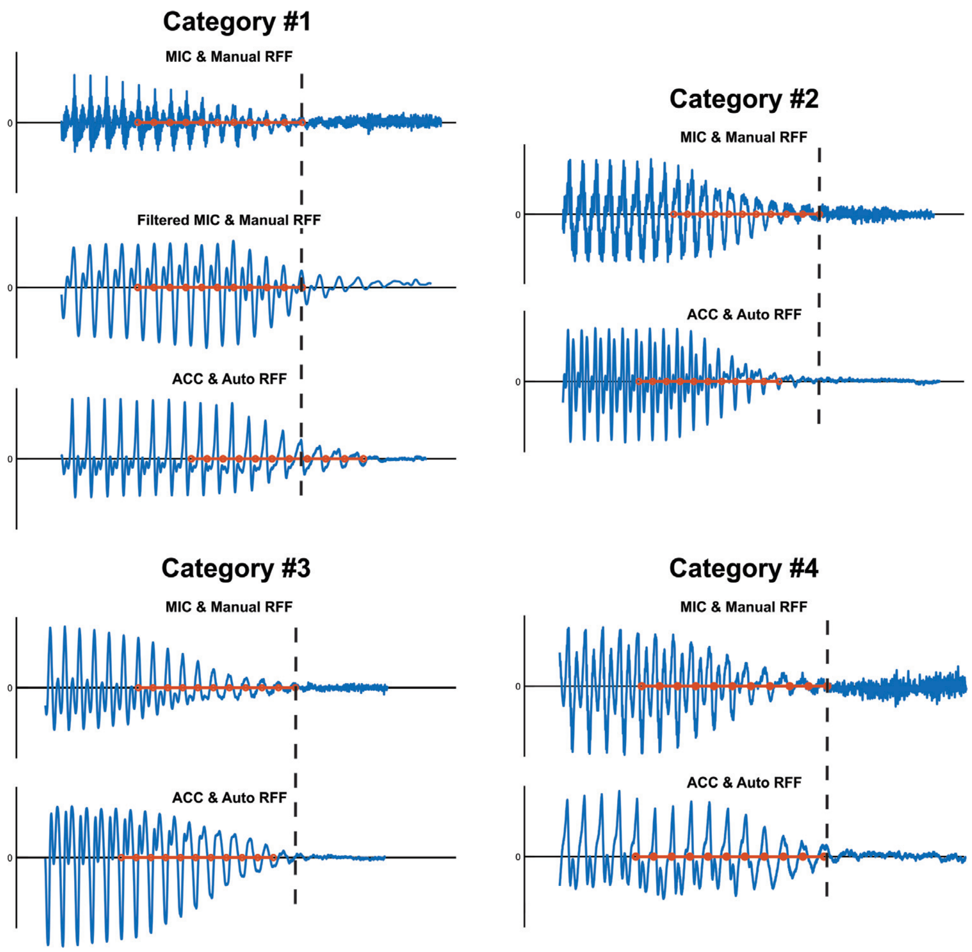

FIGURE 4. Examples of instances in which the root mean square error (RMSE) between manual RFF and automated RFF is greater than 1.0 ST for each of the four categories. The signal is plotted in blue for both microphone and accelerometer signals, with the corresponding RFF locations plotted in red. The x-axis is in arbitrary time units, and the y-axis is in arbitrary voltage units. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

the microphone signal. This is consistent with previous work that showed that manual RFF estimates from filtered microphone signals trend toward manual RFF estimates from accelerometer signals.[34] Of the 88 offset utterances that had RMSE values greater than 1.0 ST, 31 (35%) were classified as category 1. Of the 111 onset utterances that had RMSE values greater than 1.0 ST, 37 (33%) were classified as category 1.

Category 2: Utterances in which the accelerometer signal was too noisy or the final voicing cycles were not visible resulted in automatic RFF estimates failing to identify the voicing cycles closest to the fricative. This occurred for 22 (25%) offset instances and 13 (12%) onset instances.

Category 3: There were several utterances in which the ACC-refined algorithms incorrectly identified the boundary cycle despite both the microphone and accelerometer signal showing a clear voicing boundary. These errors were considered to be due to algorithmic errors, instead of signal characteristics. There were 15 (17%) offset utterances and 18 (16%) onset utterances that resulted from algorithmic errors.

Category 4: Finally, there were many utterances in which the ACC-refined algorithms correctly identified the location of the boundary cycle, but the RFF values calculated from the microphone and accelerometer signals were still markedly different. In these utterances, the automated algorithms identified the same boundary location as the trained technicians. Thus, remaining errors were a result of differences in the f_o contour estimates. These differences could either be attributed to algorithmic differences between contour estimation algorithms (ie, autocorrelation in Praat for manual RFF vs. Halcyon in automated RFF), or innate differences in the content of the two signals. There were 20 (23%) offset instances and 43 (39%) onset instances due to f_o contour errors.

DISCUSSION

Performance of Algorithms

The current ACC-refined algorithms resulted in comparable errors to both the MIC-original and MIC-refined algorithms in a novel test set. All three versions of the algorithm demonstrated similar average Pearson's correlation coefficients, with greater correlations for offset 10 than for onset 1. An average MBE of 0.03 was smallest in the ACC-refined algorithms, whereas an average RMSE of 0.30 was identical to the MIC-original algorithms and slightly higher than the MIC-refined algorithms. When inspecting errors at the cycles closest to the fricative, ACC-refined had a smaller MBE and RMSE for offset 10 when compared to the MIC-original algorithms, but larger errors when compared to the MIC-refined algorithms. ACC-refined algorithms had a larger MBE and RMSE for onset 1 when compared to both the MIC-original and MIC-refined algorithms.

The RMSE values were notably larger than the corresponding MBE values, as seen in previous iterations of the algorithms.[16,17] Whereas MBE is the result of averaging directional errors across multiple RFF instances, RMSE is the result of averaging the error magnitudes of each individual RFF instance. This indicates that although estimates of individual RFF utterances may be inaccurate, the average estimate from multiple RFF utterances results in a relatively low error when compared to the average manual estimate. As a result, much like previous iterations of the algorithms, the current ACC-refined algorithms should be used to calculate average RFF estimates across multiple utterances from the same speaker.

The current ACC-refined algorithms resulted in smaller errors for offset cycles than for onset cycles. This is not surprising given that voicing onset is more abrupt than voicing offset, resulting in faster changing RFF values.[10] Thus, if there is an error in the location of the boundary cycle approximation, it would be expected that this would result in a greater error in the RFF value in onset cycles than in offset cycles.

Error Analysis

Post hoc analysis of utterances in which the RMSE was larger than 1.0 ST revealed that many of these errors could be attributed to a mismatch in the f_o contours (Category 4; 23% of Offset instances and 39% of Onset instances). Halcyon uses normalized cross-correlation which could result in inherently different cycle definitions than the autocorrelation performed in Praat during manual RFF estimation. Furthermore, manual RFF allows the technician to identify cycles by either peaks or troughs, which could result in notably different RFF than the cycles defined by Halcyon. This could imply that errors in Category 4 are caused by simply using different f_o estimation methods even when the boundary cycle location was correctly identified. However, it could also indicate that there is an innate difference in the information between the microphone and the accelerometer signals.

In order to further investigate the cause of the error when the boundary cycle was correctly identified, author M.D.G. completed manual RFF estimates on the accelerometer signals for each of the utterances classified as category 4. Manual RFF was completed using Praat (ie, using the same f_o estimation method as the manual microphone-based RFF estimates), but the boundary cycle was selected to be identical to the boundary cycle identified by the ACC-refined algorithms for each instance. If these new manual accelerometer-based RFF values were similar to the manual microphone-based RFF values, then the error between the manual and automatic RFF values would likely be due to differences in the f_o estimation method. If, however, these values were still different, then the error was likely a result of differences in the information presented in the microphone and accelerometer signal.

Manual accelerometer-based RFF was calculated for 20 offset utterances and 43 onset utterances in which the boundary cycle was correctly time-aligned in the automated RFF estimate. The average RMSE between the manual accelerometer-based RFF estimates and the manual microphone-based RFF estimates were calculated for offset 10 and onset 1 as 1.37 and 1.30 ST, respectively. In comparison, the average RMSE between the automated accelerometer-based RFF estimates and the manual microphone-based RFF estimates were calculated for offset 10 and onset 1 as 1.72 and 1.90 ST, respectively. Thus, even when using an identical f_o estimation method via manual estimation, RFF errors were only reduced by 20% to 32%, indicating that there were innate differences between the information presented in the microphone and accelerometer signals that the current ACC-refined algorithms could not account for.

The result that there were errors based on inherent differences between the two signals was not surprising. One study compared manual RFF estimates derived from microphone and accelerometer signals and found that signal type had a small, but significant effect on manual RFF estimates.[38] The authors reasoned that this effect was due to differences caused by what is captured in each signal. Whereas the accelerometer signals capture neck-surface vibration resulting from laryngeal acoustics and vocal fold collision transmitted through subglottal resonances and neck tissue, the microphone signals capture acoustic information ie, also affected by the vocal tract shape, movement of articulators, environmental noise, and radiation characteristics of the mouth. Additionally, microphone signals are more subject to masking noise from coarticulation, which can mask additional voicing cycles during manual estimation. Based on the post hoc analysis of large error instances, coarticulation masking resulted in inaccurate boundary identification in 35% of offset instances and 33% of onset instances (Category 1). Thus, there are inherent errors when comparing the RFF values of two different types of signals, as expected based on previous studies in manual RFF.[34,38] This may explain why the semi-automated MIC-refined algorithms were able to achieve lower RFF errors than the ACC-refined algorithms. Even if both the MIC-refined and the ACC-refined algorithms successfully identify the proper boundary cycle, automated microphone-based RFF estimates will be inherently closer to the manual RFF estimates from the same type of signal.

Potential for EMA and Ambulatory Monitoring

Although the current ACC-refined algorithms resulted in somewhat larger errors than the MIC-refined algorithms, the resulting MBE errors were still small when considering the intended purpose of the ACC-refined algorithms. Specifically, ambulatory monitoring may be used to longitudinally monitor the vocal function of hyperfunctional individuals throughout voice therapy.[24] In a study that calculated RFF values in individuals with vocal hyperfunction prior to and following voice therapy, Stepp et al observed that RFF values, on average, increased by 0.50 ST for offset cycle 10 and 0.81 ST for onset cycle 1 following voice therapy.[10] RMSE values were large (0.65 ST and 1.07 ST for offset cycle 10 and onset cycle 1, respectively), but both MBE values (0.15 ST and −0.25 ST for offset cycle 10 and onset cycle 1, respectively) were smaller than the anticipated therapy effects, suggesting that these therapy effects are unlikely to be masked by algorithmic error when RFF estimates are averaged across multiple utterances.

Ambulatory monitoring may also be used to identify daily hyperfunctional behaviors.[24] Several studies have observed a decrease in RFF values when healthy participants were instructed to speak with increased vocal effort.[12,13,45] Vocal effort is thought to be a hallmark of hyperfunction behavior. McKenna and colleagues reported that, when compared to a typical speaking effort, individuals speaking with maximum effort have an average decrease in RFF of 0.99 ST for offset cycle 10 and 0.45 ST for onset cycle 1.[45] These changes in RFF as a result of hyperfunctional behavior are larger than the observed MBE values in the overall test set and larger than the observed MBE value for offset cycle 10 in the subset of individuals with voice disorders. This suggests that algorithmic errors will not prevent the identification of hyperfunctional behaviors via EMA and ambulatory monitoring.

Given that offset cycle 10 resulted in smaller MBE values than onset cycle 1, specifically when comparing offset 10 to onset 1, it is clear that voicing offset RFF values were more robust to algorithmic estimation errors. Both offset and onset RFF have demonstrated variable sensitivity to vocal function in different studies. One study found that an increase in RFF of offset cycle 10 was seen in 81% of individuals with vocal hyperfunction who successfully completed voice therapy.[10] In comparison, an increase in RFF of onset cycle 1 was seen in 94% of the same individuals. Another study found that RFF of offset cycle 10 was a significant predictor of listener-perception of vocal effort in individuals with healthy voices who self-modulated their vocal effort, but that RFF of onset cycle 1 was not.[13] Based on the current results, we recommend that clinical research using the current ACC-refined algorithms for EMA and ambulatory monitoring should focus on offset values for observing and monitoring vocal hyperfunction in order to reduce the impact of known estimation errors. Future research should focus on reducing the error in onset value estimates.

Limitations and Conditions for Application

The subset of speech samples used in the current study consisted of speech samples from both individuals with healthy voices and individuals with voice disorders. This distribution was chosen in order to generalize performance across a wide-range of speech samples. However, device performance may vary between individuals with and without voice disorders. Due to conflating factors such a signal quality (a majority of speech samples from individuals with voice disorders are recorded in a noisier location), this comparison is beyond the scope of the current study. Thus, future studies should investigate device performance in individuals with and without voice disorders by better controlling for confounding factors.

The current ACC-refined algorithms make several assumptions that should be considered. Unlike previous RFF algorithms, RFF estimation in the current algorithms is fully automated by automatically determining the location of fricatives in the speech sample. In order to do this, the algorithms assume three isolated VCV utterances. Though algorithmic parameters can be easily modified to accommodate a larger number of VCV utterances in each sample, the current algorithms are not equipped to automatically calculate RFF from stimuli that are not in the form of isolated VCV utterances, such as running speech. Thus, the current algorithms only function for EMA where participants would be required to specifically produce VCV utterances. Future work should focus on adapting the current algorithms for running speech in order to allow for true ambulatory monitoring of everyday voice use in natural discourse level contexts.

During fricative identification, the ACC-refined algorithms also removed all samples that did not exceed a set vowel-to-noise ratio in order to avoid samples in which the fricative locations were incorrectly identified, as well as samples that would be considered too noisy to properly identify RFF values manually. Although this threshold is a normalized ratio and was based on a wide range of samples from both healthy and disordered speakers, it is possible that this threshold is not appropriate for all accelerometer signals. Indeed, all data were collected using a Knowles accelerometer. It is important to acknowledge that the current ACC-refined algorithms were developed to optimize performance on signals acquired with specific equipment and that performance may change across different experimental set-ups.

With *a priori* knowledge that the intended purpose of the algorithms would be for EMA, design of the current ACC-refined algorithms made the assumption that there was a large number of speech samples available for each participant. Therefore, thresholds based on the vowel-to-noise ratio of the signal (see *Fricative Identification* in Methods) were designed to reject a greater number of utterances for the sake of more accurate RFF estimations. These thresholds were in addition to the rejection criteria in place for all three versions of the algorithms in which utterances are rejected if RFF patterns are not consistent with physiologically valid productions (see *Final RFF Calculation* in Methods).

Stricter rejection criteria are important to consider for applications in which the amount of data is more limited. The current dataset had a total of nine utterances per participant. After rejection, many participants only had a few usable utterances, with 5.0 and 4.7 usable utterances for offset and onset RFF, respectively, when averaging across participants with a nonzero number of usable utterances. These averages are similar to previous versions of the algorithms (5.5 and 5.3 for offset and onset in MIC-original; 4.4 and 4.3 for offset and onset in MIC-refined).[16,17] Though it is possible that RFF estimates may change with additional utterances, one study demonstrated that access to at least six utterances (prior to rejection) resulted in a stable level of association between the RFF of onset cycle 1 and the perception of vocal effort in individuals with laryngeal dystonia.[8] Therefore, use of the ACC-refined algorithms will be most appropriate when a large number of utterances (no less than six) per participant can be collected.

CONCLUSION

A set of fully automated ACC-refined algorithms was developed to calculate clinically relevant RFF estimates from accelerometer signals. When compared to the gold-standard of manual RFF estimates from microphone signals, automated RFF values from neck-placed accelerometer signals have an average MBE of 0.03 ST, with an MBE of 0.15 ST for offset 10 and an MBE of -0.25 ST for onset 1. These errors are smaller than the expected differences in RFF values following successful voice therapy for individuals with vocal hyperfunction,[10] indicating that the current algorithms could be used for EMA and ambulatory monitoring via neck-surface accelerometer signals.

APPENDIX

The following paragraphs provide supplemental detail about specific thresholds and signal processing used to identify fricative locations and voicing boundary cycle locations in the ACC-refined algorithms. This appendix, the main text, and code available online from previous versions of the algorithms[3], should provide sufficient detail to replicate this work.

Fricative Identification

The ACC-refined algorithms first calculate the location of the three fricatives in a speech sample consisting of three VCV utterances. This process begins by first identifying the vowels on either side of each fricative (six vowel segments in total). The raw accelerometer signal is first band-pass filtered with a fifth-order Butterworth filter between 100 and 1000 Hertz. The root mean square of this filtered signal is calculated over a 300-point window length with a 10-point step size. The natural logarithm of the normalized RMS is then smoothed using a 500-point moving average filter.

Each point in the signal is then discretized into three values (0, 1, 2) using the standard MATLAB function "discretize." Points in the signal classified as the largest in magnitude (a value of 2) are considered vowel segments, whereas all other points (values of 0 and 1) are considered fricative or silence segments. All vowel segments less than 90 points in length are considered to be misclassified and are given a value of 0. If there are less than six vowel segments, then segments of the signal that are classified as the second largest in magnitude (a value of 1) are added as additional vowel segments. Vowel segments less than 90 points in length are again given a value of 0.

Following this first iteration, if there are still less than six vowel segments, the signal will be rediscretized into four values (0, 1, 2, and 3), with the same process being repeated (ie, points with a value of 3 are assigned as vowel segments, followed by points with a value of 2 if six vowel segments of proper length are not identified, followed by points with a value of one if six vowel segments of proper length are still not identified). If, after all three nonzero values are added as additional vowel segments, there are still less than six vowel segments of proper length, the signal is again rediscretized using five values. Following this, the sample is rejected if six vowel segments of proper length are not located.

If six or more vowel segments are successfully identified, the fricative locations are then identified as the median point between each pair of the six largest vowel segments. These fricative locations are used to identify voicing offset and onset for each VCV utterance. An example of this discretization is shown in Figure 5.

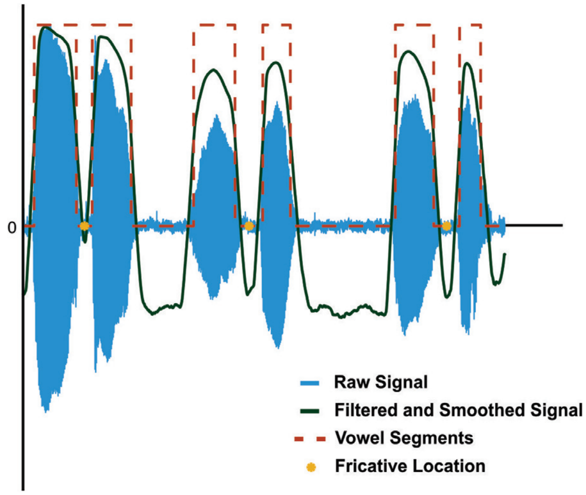

FIGURE 5. Example of fricative identification in a speech sample of three utterances of /ufu/, in which the blue line is the raw accelerometer signal, the green line is the filtered and smoothed version of the accelerometer, the orange line shows the vowel segments, and the yellow circles correspond to the fricative locations. The x-axis is arbitrary time units and the y-axis is arbitrary voltage units.(For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

[3] The semi-automated RFF estimation algorithms for microphone signals can be downloaded from http://sites.bu.edu/stepplab/research/rff/.

Fundamental Frequency (f_o) Estimation

Estimation of the fundamental frequency (f_o) contour was performed using a custom-made MATLAB script that implemented the Halcyon model[43] for f_o estimation. No changes were made to the model described by Azarov et al.

Offset Boundary Cycle Identification

All information pertaining to the identification of the offset boundary cycle location is provided in the main text. No additional information is needed.

Onset Boundary Cycle Identification

An onset boundary cycle was determined by first finding one boundary cycle location using the same method as described for offset boundary cycle identification, henceforth referred to as the log likelihood boundary cycle. A second boundary location was calculated based on z-scores and conditional thresholding. The reason for identifying two boundary locations is discussed in the main text. The focus of the following text will be on the methodology.

Unlike in the log likelihood boundary cycle identification, the normalized peak-to-peak amplitude was calculated on a cycle-by-cycle basis such that a vector of values for each feature was calculated based on the cycle locations from the f_o contour. The mean and standard deviation of the first five values in the peak-to-peak vector (corresponding to the cycles closest to the fricative) were calculated and used to normalize each peak-to-peak value in the vector into a z-score using the formula: z-score = (x − μ) / σ, in which x is the given value, μ is the mean, and σ is the standard deviation.

Analyzing the peak-to-peak z-score patterns of individual utterances showed that there were a range of possible patterns. As a result, conditional thresholding was implemented to target specific patterns. These conditional thresholds were derived from systematically changing each threshold until the greatest number of boundary samples could be correctly selected in the training set. The optimized thresholding is as follows. First, from the vector of peak-to-peak z-scores, the cycle immediately preceding the first cycle to have a z-score over 19 was marked as the boundary cycle. However, if the peak-to-peak z-score of at least one cycle was above 50 and the peak-to-peak z-score of the last cycle in the vector was also above 50, the boundary cycle would be reselected as the last cycle to have a z-score below 19. Similarly, if there was at least one cycle with a peak-to-peak z-score above 50, but the peak-to-peak z-score of the last cycle in the vector was not above 50, then the peak-to-peak z-score vector was normalized and the boundary cycle was set as the last cycle to have a normalized peak-to-peak z-score below 0.16. Finally, there were many onset utterances in which the maximum peak-to-peak z-score was much larger than the majority of the utterances in the training set. As a result, if the peak-to-peak z-score of at least one cycle was above 220, then the peak-to-peak z-score vector was normalized and the boundary cycle was set as the last cycle to have a normalized z-score below 0.16. Without these conditional thresholds in place, the first offset cycle or the last onset cycle would often be incorrectly identified as the voicing boundary cycle, resulting in large RFF errors. The current threshold values resulted in the greatest correspondence between the automatically identified boundary cycles and the manually identified boundary cycles.

The final boundary cycle from this conditional thresholding was considered the z-score boundary cycle. The median of the z-score boundary cycle and the log likelihood ratio boundary cycle was calculated and rounded up to the nearest cycle resulting in a final onset boundary cycle.

REFERENCES

1. Baken R, Orlikoff RF. The effect of articulation on fundamental frequency in singers and speakers. *J Voice*. 1987;1:68–76.
2. Halle M. On distinctive features and their articulatory implementation. *Nat Lang Linguistic Theory*. 1983;1:91–105.
3. Lofqvist A, Baer T, McGarr N, Story R. The cricothyroid muscle in voicing control. *J Acoust Soc Am*. 1989;85:1314–1321.
4. Löfqvist A, Koenig LL, McGowan RS. Vocal tract aerodynamics in/aCa/utterances: measurements. *Speech Commun*. 1995;16:49–66.
5. Fukui N, Hirose H. Laryngeal adjustments in Danish voiceless obstruent production. annual bulletin. *Res Institute Logop Phoniatr*. 1983;17:61–71.
6. Stepp CE. Relative fundamental frequency during vocal onset and offset in older speakers with and without Parkinson's disease. *J Acoust Soc Am*. 2013;133:1637–1643.
7. Goberman AM, Blomgren M. Fundamental frequency change during offset and onset of voicing in individuals with Parkinson disease. *J Voice*. 2008;22:178–191.
8. Eadie TL, Stepp CE. Acoustic correlate of vocal effort in spasmodic dysphonia. *Ann Otol Rhinol Laryngol*. 2013;122:169–176.
9. Stepp CE, Hillman RE, Heaton JT. The impact of vocal hyperfunction on relative fundamental frequency during voicing offset and onset. *J Speech Lang Hear Res*. 2010;53:1220–1226.
10. Stepp CE, Merchant GR, Heaton JT, Hillman RE. Effects of voice therapy on relative fundamental frequency during voicing offset and onset in patients with vocal hyperfunction. *J Speech Lang Hear Res*. 2011;54:1260–1266.
11. Stepp CE, Sawin DE, Eadie TL. The relationship between perception of vocal effort and relative fundamental frequency during voicing offset and onset. *J Speech Lang Hear Res*. 2012;55:1887–1896.
12. Lien YA, Michener CM, Eadie TL, Stepp CE. Individual monitoring of vocal effort with relative fundamental frequency: relationships with aerodynamics and listener perception. *J Speech Lang Hear Res*. 2015;58:566–575.
13. McKenna VS, Stepp CE. The relationship between acoustical and perceptual measures of vocal effort. *J Acoust Soc Am*. 2018;144:1643.
14. Kempster GB, Gerratt BR, Verdolini Abbott K, Barkmeier-Kraemer J, Hillman RE. Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol. *Am J Speech Lang Pathol*. 2009;18:124–132.
15. Boersma P. Praat, a system for doing phonetics by computer. *Glot International*. 2001;5:341–345.
16. Lien Y-AS, Heller Murray ES, Calabrese CR, et al. Validation of an algorithm for semi-automated estimation of voice relative fundamental frequency. *Ann Otol Rhinol Laryngol*. 2017;126:712–716.
17. Vojtech JM, Segina RK, Buckley DP, et al. Refining algorithmic estimation of relative fundamental frequency: accounting for sample characteristics and fundamental frequency estimation method. *J Acoust Soc Am*. 2019;146:3184.
18. Camacho A. On the use of auditory models' elements to enhance a sawtooth waveform inspired pitch estimator on telephone-quality signals. *2012 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA)*. 2012.

19. Camacho A. *SWIPE: A Sawtooth Waveform Inspired Pitch Estimator for Speech and Music*. University of Florida Gainesville; 2007.
20. Camacho A, Harris JG. A sawtooth waveform inspired pitch estimator for speech and music. *J Acoust Soc Am*. 2008;124:1638–1652.
21. Robb MP, Smith AB. Fundamental frequency onset and offset behavior. *J Speech Lang Hear Res*. 2002.
22. Cortes JP, Espinoza VM, Ghassemi M, et al. Ambulatory assessment of phonotraumatic vocal hyperfunction using glottal airflow measures estimated from neck-surface acceleration. *PLoS One*. 2018;13:e0209017.
23. Mehta DD, Van Stan JH, Hillman RE. Relationships between vocal function measures derived from an acoustic microphone and a subglottal neck-surface accelerometer. IEEE/ACM transactions on audio. *IEEE/ACM Trans Audio Speech Lang Process,*. 2016;24:659–668.
24. Mehta DD, Van Stan JH, Zanartu M, et al. Using ambulatory voice monitoring to investigate common voice disorders: research update. *Front Bioeng*. 2015;3:155.
25. Ortiz AJ, Toles LE, Marks KL, et al. Automatic speech and singing classification in ambulatory recordings for normal and disordered voices. *J Acoust Soc Am*. 2019;146:EL22.
26. Van Stan JH, Mehta DD, Sternad D, Petit R, Hillman RE. Ambulatory voice biofeedback: relative frequency and summary feedback effects on performance and retention of reduced vocal intensity in the daily lives of participants with normal voices. *J Speech Lang Hear Res*. 2017;60:853–864.
27. Popolo PS, Svec JG, Titze IR. Adaptation of a Pocket PC for use as a wearable voice dosimeter. *J Speech Lang Hear Res*. 2005;48:780–791.
28. Švec JG, Titze IR, Popolo PS. Estimation of sound pressure levels of voiced speech from skin vibration of the neck. *J Acoust Soc Am*. 2005;117:1386–1394.
29. Titze IR, Švec JG, Popolo PS. Vocal dose measures: quantifying accumulated vibration exposure in vocal fold tissues. *J Speech Lang Hear Res*. 2003;46:919–932.
30. Hunter EJ. Teacher response to ambulatory monitoring of voice. *Logoped Phoniatr Vocol*. 2012;37:133–135.
31. Zanartu M, Ho JC, Kraman SS, Pasterkamp H, Huber JE, Wodicka GR. Air-borne and tissue-borne sensitivities of bioacoustic sensors used on the skin surface. *IEEE Trans Biomed Eng*. 2009;56:443–451.
32. Cheyne HA, Hanson HM, Genereux RP, Stevens KN, Hillman RE. Development and testing of a portable vocal accumulator. *J Speech Lang Hear Res*. 2003.
33. Zañartu M, Ho JC, Mehta DD, Hillman RE, Wodicka GR. Subglottal impedance-based. *IEEE Trans Audio Speech Lang Process*. 2013;21:1929–1939.
34. Lien Y-AS, Stepp CE. Comparison of voice relative fundamental frequency estimates derived from an accelerometer signal and low-pass filtered and unprocessed microphone signals. *J Acoust Soc Am*. 2014;135:2977–2985.
35. Sugimoto T, Hiki S. Extraction of the pitch of a voice from the vibration of the outer skin of the trachea. *J Acoust Soc Japan*. 1960;16:291–293.
36. Mehta DD, Espinoza VM, Van Stan JH, Zanartu M, Hillman RE. The difference between first and second harmonic amplitudes correlates between glottal airflow and neck-surface accelerometer signals during phonation. *J Acoust Soc Am*. 2019;145:EL386.
37. Van Stan JH, Mehta DD, Ortiz AJ, et al. Differences in weeklong ambulatory vocal behavior between female patients with phonotraumatic lesions and matched controls. *J Speech Lang Hear Res*. 2020;63:372–384.
38. Lien Y-AS, Calabrese CR, Michener CM, et al. Voice relative fundamental frequency via neck-skin acceleration in individuals with voice disorders. *J Speech Lang Hear Res*. 2015;58:1482–1487.
39. Heller Murray ES, Lien YS, Van Stan JH, et al. Relative fundamental frequency distinguishes between phonotraumatic and non-phonotraumatic vocal hyperfunction. *J Speech Lang Hear Res*. 2017;60:1507–1515.
40. Bayo MV, Garcia AM, Garcia A. Noise levels in an urban hospital and workers' subjective responses. *Arch Environ Health*. 1995;50:247–251.
41. Patel RR, Awan SN, Barkmeier-Kraemer J, et al. Recommended protocols for instrumental assessment of voice: American speech-language-hearing association expert panel to develop a protocol for instrumental assessment of vocal function. *Am J Speech-Lang Pathol*. 2018;27:887–905.
42. Lien YA, Gattuccio CI, Stepp CE. Effects of phonetic context on relative fundamental frequency. *J Speech Lang Hear Res*. 2014;57:1259–1267.
43. Azarov E, Vashkevich M, Petrovsky A. Instantaneous pitch estimation algorithm based on multirate sampling. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2016.
44. McKenna VS, Heller Murray ES, Lien YS, Stepp CE. The relationship between relative fundamental frequency and a kinematic estimate of laryngeal stiffness in healthy adults. *J Speech Lang Hear Res*. 2016;59:1283–1294.
45. McKenna VS, Diaz-Cadiz ME, Shembel AC, Enos NM, Stepp CE. The relationship between physiological mechanisms and the self-perception of vocal effort. *J Speech Lang Hear Res*. 2019;62:815–834.