JSLHR

Research Article

Adductory Vocal Fold Kinematic Trajectories During Conventional Versus High-Speed Videoendoscopy

Manuel Diaz-Cadiz,[a] Victoria S. McKenna,[a] Jennifer M. Vojtech,[a,b] and Cara E. Stepp[a,b,c]

Objective: Prephonatory vocal fold angle trajectories may supply useful information about the laryngeal system but were examined in previous studies using sigmoidal curves fit to data collected at 30 frames per second (fps). Here, high-speed videoendoscopy (HSV) was used to investigate the impacts of video frame rate and sigmoidal fitting strategy on vocal fold adductory patterns for voicing onsets.
Method: Twenty-five participants with healthy voices performed /ifi/ sequences under flexible nasendoscopy at 1,000 fps. Glottic angles were extracted during adduction for voicing onset; resulting vocal fold trajectories (i.e., changes in glottic angle over time) were down-sampled to simulate different frame rate conditions (30–1,000 fps). Vocal fold adduction data were fit with asymmetric sigmoids using 5 fitting strategies with varying parameter restrictions. Adduction trajectories and maximum adduction velocities were compared between the fits and the actual HSV data. Adduction trajectory errors between HSV data and fits were evaluated using root-mean-square error and maximum angular velocity error.
Results: Simulated data were generally well fit by sigmoid models; however, when compared to the actual 1,000-fps data, sigmoid fits were found to overestimate maximum angle velocities. Errors decreased as frame rate increased, reaching a plateau by 120 fps.
Conclusion: In healthy adults, vocal fold kinematic behavior during adduction is generally sigmoidal, although such fits can produce substantial errors when data are acquired at frame rates lower than 120 fps.

Transnasal flexible laryngoscopy is a staple in clinical practice for investigating voice production, enabling visualization of laryngeal anatomy and physiology. However, clinical examination of the vocal structures using traditional flexible laryngoscopy is hindered by the standard time resolution of the camera used to record endoscopic images. Specifically, clinical laryngoscopic images are typically recorded at 30 frames per second (fps), which is far too slow to resolve the vibratory dynamics of the vocal folds (VFs; Aghdam et al., 2017; Ishii et al., 2011). As such, traditional flexible laryngoscopy has also been applied to examine gross abductory and adductory gestures. These movements, although over an order of magnitude slower than VF vibrations, are still quite fast: Average maximal abductory and adductory movement times occur within 104–227 ms (Dailey et al., 2005). Therefore, the standard 30-fps sampling rate can only measure an average of three to seven frames during each abductory or adductory movement. Due to this crude time resolution and thus limited data, VF kinematic features during abductory and adductory gestures have primarily been examined by manually estimating VF (glottic) angles extending from the anterior commissure, posterior to the VF processes. The sparse data have been subsequently fitted with a cubic (Dailey et al., 2005) or, more recently, an asymmetric sigmoidal function (Britton et al., 2012; McKenna, Heller Murray, Lien, & Stepp, 2016; Stepp, Hillman, & Heaton, 2010). These fits allow for the estimation of features of the movement trajectory (e.g., maximum velocity), though the accuracy of these techniques has yet to be fully vetted.

Recently, the incorporation of new imaging technology in medical instrumentation has led to improvements in time resolution of laryngoscopy procedures (Mehta & Hillman, 2012). Laryngeal examination via high-speed videoendoscopy (HSV) provides higher temporal resolution (much faster than 30 fps), which allows for accurate estimation of both

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

Editor-in-Chief: Julie Liss
Editor: Jack Jiang
Received October 5, 2018
Revision received December 30, 2018
Accepted February 11, 2019
https://doi.org/10.1044/2019_JSLHR-S-18-0405

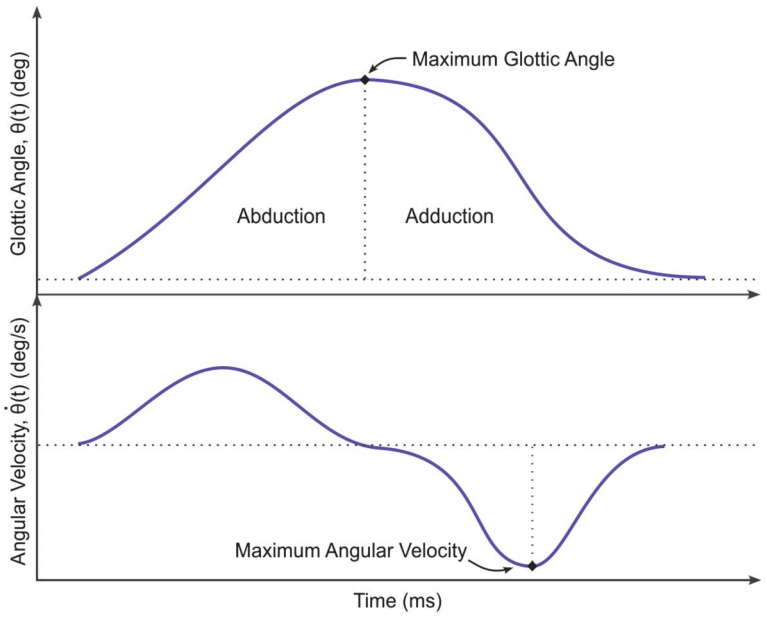
Disclosure: *The authors have declared that no competing interests existed at the time of publication.*

vibratory features (Deliyski, Powell, Zacharias, Gerlach, & de Alarcon, 2015; Patel, Dubrovskiy, & Döllinger, 2014) and VF kinematic trajectories (Freeman, Woo, Saxman, & Murry, 2012; Iwahashi, Ogawa, Hosokawa, Kato, & Inohara, 2016). Several approaches using HSV have been developed to examine characteristics of steady-state phonation and associated vocal onsets and offsets (e.g., Aghdam et al., 2017; Braunschweig, Flaschka, Schelhorn-Neise, & Döllinger, 2008; Döllinger, Dubrovskiy, & Patel, 2012; Freeman et al., 2012; Guzman et al., 2017; Ikuma, Kunduk, Fink, & McWhorter, 2016; Iwahashi et al., 2016; Kunduk, Döllinger, McWhorter, & Lohscheller, 2010; Kunduk, Vansant, Ikuma, & McWhorter, 2017; Kunduk, Yan, McWhorter, & Bless, 2006; Mehta, Deliyski, Quatieri, & Hillman, 2011; Patel et al., 2014; Patel, Forrest, & Hedges, 2017; Patel, Unnikrishnan, & Donohue, 2016; Patel, Walker, & Sivasankar, 2016; Watanabe, Kaneko, Sakaguchi, & Takahashi, 2016; Woo, 2017; Yamauchi et al., 2016; Zacharias, Deliyski, & Gerlach, 2018); yet, to date, there are few studies characterizing VF kinematics of laryngeal posturing during connected speech.

VF kinematic features, particularly adductory movements, are of interest due to previous work, suggesting that they may reveal information about biomechanical characteristics of the laryngeal system (Cooke, Ludlow, Hallett, & Scott Selbie, 1997). A schematic of the expected behavior of glottic angle kinematics during the transition from a vowel to a voiceless consonant and back to a vowel is illustrated in Figure 1. Stepp et al. (2010) used glottic angle measurements, which were introduced by Dailey et al. (2005), to quantify adduction trajectories. In particular, an estimate of laryngeal kinematic stiffness was calculated by normalizing the maximum angular velocity (MAV) of the VFs during adduction by the maximum extent of the glottic angle during movement. The authors implemented a simple virtual trajectory model of VF kinematics to determine a relationship between changes in intrinsic laryngeal muscle stiffness parameters in the model and the kinematic stiffness ratios of the resulting simulated movement. Results showed that increases in stiffness were associated with changes in kinematic stiffness ratios, further supporting the use of VF kinematics as an estimate of laryngeal stiffness.

Although adductory kinematics have shown promise as indicators of laryngeal function, the VF adductory kinematics were acquired using conventional speed videoendoscopy systems at a frame rate of 30 fps (Britton et al., 2014, 2012; Dailey et al., 2005; McKenna et al., 2016; Stepp et al., 2010). Due to restrictions in frame rate, many researchers reconstructed adduction trajectories by fitting a parametric model to estimate adduction velocity (Britton et al., 2014, 2012; Dailey et al., 2005; McKenna et al., 2016; Stepp et al., 2010). Methodological decisions regarding the appropriate type of fitting model have been described as a possible source of measurement error in the resulting estimates of stiffness. Indeed, researchers using HSV to investigate steady-state voice production have shown that a frame rate of 4,000 fps is required for an accurate assessment of VF vibratory features (Deliyski et al., 2015). Recently, a study by Iwahashi et al. (2016) used an HSV system recording at 4,000 fps to investigate whether the parametric fit approach used in previous studies was a valid means of analyzing laryngeal kinematics. The authors focused on examining motion of the laryngeal structures

Figure 1. Schematic of glottic angle kinematics (top: glottic angle, bottom: glottic angular velocity) during voicing offset and onset surrounding a voiceless consonant.



during sustained phonation and throat clears. Although the adduction trajectories showed polynomial-like curves rather than sigmoidal curves during throat clears, the authors suggested that adduction trajectories were generally well fit by a sigmoidal model for sustained phonation (Iwahashi et al., 2016). However, these findings were based on qualitative impressions from sustained phonation. As such, it is unclear as to whether these findings are generalizable to continuous speech and whether temporal resolution and fitting strategy impact the appropriateness of sigmoidal fits to VF adductory kinematics.

Objectives

The purpose of this study was to (a) determine whether VF adductory behaviors during voicing onset follow a sigmoidal trajectory, as suggested by Stepp et al. (2010); (b) determine whether sampling rates affect the appropriateness of the sigmoidal fit to VF trajectories; and (c) identify the strategies that produce the most appropriate model fits. We addressed these aims by simulating different frame rates and applying different fitting strategies in order to determine the errors between sigmoidal fits of simulated data and the actual adduction trajectories obtained with HSV.

Method

Participants

Twenty-five participants with healthy voices aged 18–29 years (16 women, nine men; $M = 20.8$ years, $SD = 2.9$ years) produced vowel–consonant–vowel (VCV) utterances of the nonsense word /ifi/. Participants were speakers of Standard American English with no formal or trained singing experience beyond grade school, all of whom reported no prior history of speech, language, or hearing disorders. All participants were nonsmokers and were screened for healthy vocal function by a certified speech-language pathologist via auditory–perceptual screening and examination via transnasal flexible laryngoscopy. Informed consent was obtained, and the study was carried out in compliance with the Boston University Institutional Review Board.

Participants were first trained to repeat a string of the utterance /ifi/ as two sets of four /ifi/ productions, with a pause for breath between the sets (i.e., /ifi ifi ifi ifi/, pause, /ifi ifi ifi ifi/). The VCV utterance /ifi/ creates an abductory gesture during the /f/ phoneme and an adductory revoicing gesture for the following /i/ vowel (Lien, Gattuccio, & Stepp, 2014). We chose the phoneme /i/ because it creates an open pharynx for laryngeal visualization under transnasal flexible laryngoscopy (McKenna et al., 2016). A metronome was used to train participants to modulate their vocal rate across three speeds (Hetrich & Ackermann, 1995; Ostry & Munhall, 1985): slow rate (SR; 50 words per minute [wpm]), regular rate (RR; 65 wpm), and fast rate (FR; 80 wpm). Participants were then instructed to "increase your effort during your speech as if you are trying to push your air out" while maintaining a comfortable speaking rate and volume in order to modulate their vocal effort. Specifically, participants were trained to modulate their vocal effort across three levels: mild (MIL), moderate (MOD), and maximum (MAX). Participants were trained to produce /ifi/ strings at these varying speeds and levels of effort as a means of altering laryngeal stiffness and tension, since kinematic stiffness ratios have been determined during simulated modulations in previous studies (McKenna et al., 2016; Stepp et al., 2010).

Procedure

Participants were seated for the duration of the experimental recordings. Microphone signals were recorded using a directional headset microphone (Shure SM35 XLR) placed 45° from the midline and 7 cm from the lips. A neck-surface accelerometer (BU Series 21771 from Knowles Electronic) was placed on the anterior neck, superior to the thyroid notch and inferior to the cricoid cartilage using double-sided adhesive. Microphone and accelerometer signals were preamplified (Xenyx Behringer 802 Preamplifier) and digitized at 30 kHz (National Instruments 6312 USB).

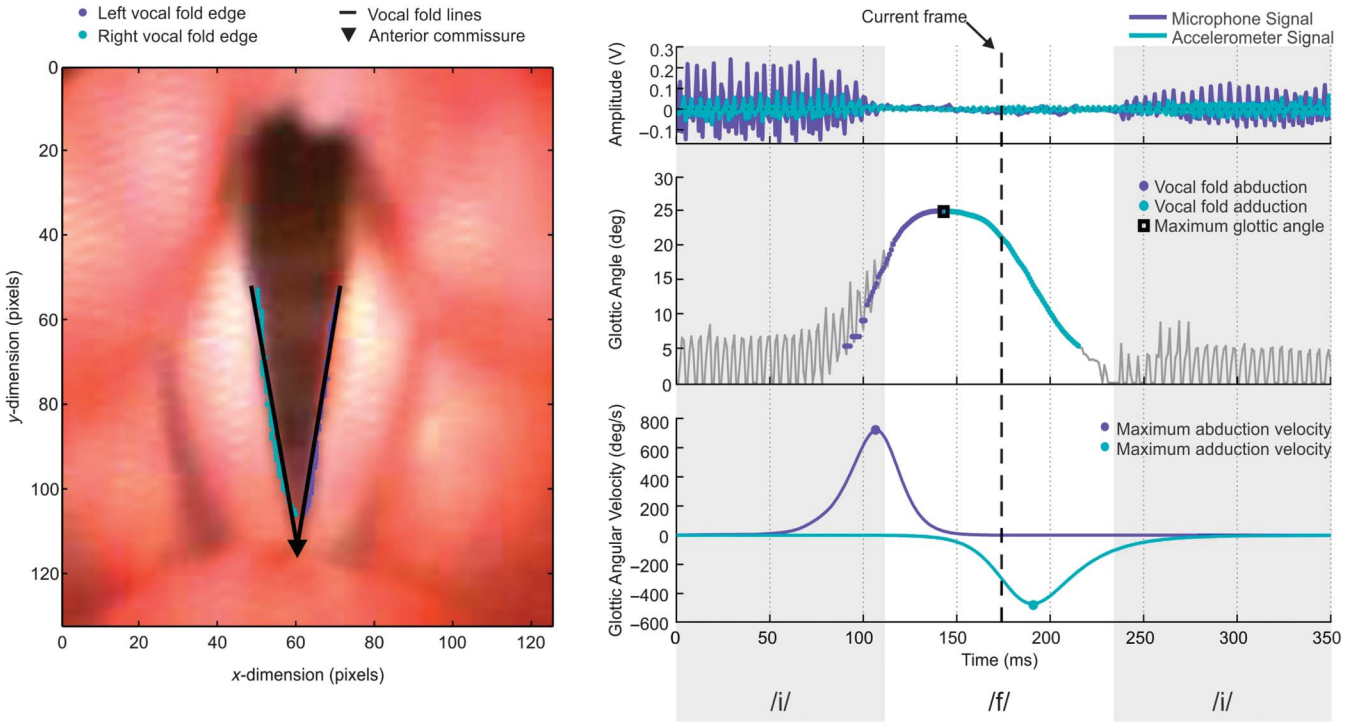
Participants underwent transnasal flexible laryngoscopy (PENTAX Medical, Model FNL-7RP3, 2.4 mm). Participants were instructed to repeat each condition twice (i.e., SR, RR, FR, MIL, MOD, MAX) while maintaining their voice intensity as best as possible. Each condition was recorded twice in order to maximize the number of VCV utterances in each condition in terms of partition recording time and camera storage capacity. The endoscope was attached to a camera (FASTCAM Mini AX100l, Model 540K-C-16GB, 256 × 256 pixels) with a 40-mm optical lens adapter and a steady xenon light source (300-W Kay-Pentax Model 7162B). Video images were acquired using Photron Fastcam Viewer software (v.3.6.6) at a frame rate of 1,000 fps. HSV image acquisition was synchronized with microphone and accelerometer recordings by using the HSV camera as the "master" (i.e., the FASTCAM Mini AX100l camera synchronization signals). Since signal synchronization happens in real time, it is expected that a natural time delay between microphone and accelerometer recordings will occur, mainly due to the distance from glottis to microphone. However, such time delays (less than 1 ms) were considered negligible in comparison with the time scale of the observed adductory motions.

Data Processing

Glottic Angle Extraction

A semiautomated technique was developed to extract glottic angles from concurrently recorded HSV, microphone, and accelerometer signals, which is fully described in the Appendix. In brief, users interacted with a graphical user interface (see Figure 2 for user display) to automatically extract glottic angles from either the microphone and accelerometer signal. If the technicians did not agree with

Figure 2. Glottic angle extraction results from a healthy participant producing the vowel–consonant–vowel utterance /ifi/ using typical vocal rate and vocal effort. The left panel shows the current frame of the high-speed video, with left and right vocal fold edges and respective fitted lines displayed. The top right plot shows the overlaid microphone and accelerometer signal. The middle and bottom plots display the algorithmic output in terms of glottic angle waveform (middle) and glottic angular velocity (bottom). The location in time of the frame shown to the left is marked with a black dotted line on the right-hand plots for reference. Shaded regions represent the duration of time corresponding to the vowel /i/ in the utterance /ifi/.



the resulting glottic angle waveform, the technicians manually marked glottic angles at a down-sampled rate of 50 Hz (see Manual Angle Marking in the Appendix for more details). Following manual angle marking, the algorithm incorporated manual angle data as a reference, which improved VF tracking in the event of camera or epiglottis motion. The result of this process was a glottic angle time waveform from each /ifi/ instance.

Using this methodology, a total of 2,053 recorded /ifi/ instances were processed. The technicians accepted initial automated angle results in 69.8% of the cases, with an additional 10.2% accepted after the aid of manual angle markings. This resulted in 1,642 /ifi/ productions that were considered viable for further analysis, with 223 corresponding to SR, 288 to RR, 294 to FR, 293 to mild effort, 276 to moderate effort, and 268 to maximum effort. Of the 20% of /ifi/ productions that were rejected, 9.9% were considered unusable due to errors in algorithm estimation of the glottic angle waveform, and the remaining 10.1% of cases were rejected prior to algorithmic analysis due to glottal occlusion by supraglottal structures.

Technician Training and Reliability

Four technicians were trained to use the algorithm to extract glottic angle data from the resulting HSV data. Prior to extraction of experimental data, raters completed a training module that consisted of extracting glottic angles from a subset of HSV data segments that included variations in scope angles, speech rate, and effort levels. An a priori intraclass correlation coefficient ICC (2, 1) $\geq$ .80 was required for a rater to be considered proficient in glottic angle extraction when using the proposed algorithm.

Adduction Trajectories and Velocities

In order to investigate whether VF adductory behavior is sigmoidal and whether sampling rate or fitting strategy affects the appropriateness of the sigmoidal fits, we simulated different frame rates and applied different fitting strategies to the glottic angle data. We then calculated MAV and root-mean-square (RMS) errors between the actual adduction trajectories obtained via HSV ("nonparametric analysis") and the sigmoidal fits of simulated data ("parametric analysis") as a function of simulated frame rate and sigmoidal fitting strategy. MAV errors were computed as a signed difference value, wherein a negative quantity signified that the magnitude of adductory velocity, as computed via a sigmoidal fit, was larger than the actual extracted trajectory.

Nonparametric analysis. Adduction trajectories and MAV estimates were obtained by directly analyzing the raw glottic angle data. Instead of fitting a model to the raw glottic angle data, the data were cropped and low-pass filtered to produce a smoothed glottal response; this response was used as a reference for adduction trajectories.

Cropping the angle data was necessary in order to remove VF oscillations captured in the raw angle trajectory response. These oscillations did not pertain to abductory and adductory movements that occur during voice offset and onset, respectively, in the VCV utterance /ifi/. Angle trajectories were cropped using an envelope-thresholding method, which consisted of overlapping high (E^{H}) and low (E^{L}) angle waveform envelopes by a factor, S, which was proportional to the average of envelope distances (d_{i}). The expression of S can be seen in Equation 1 as follows:

$$S = \frac{K}{N_D} \sum_{\mathrm{D}} d_{\mathrm{i}}, D = \left[d_{\mathrm{i}} : d_{\mathrm{i}} \le p_{10}\left(\left| E^{\mathrm{H}} - E^{\mathrm{L}} \right| \right) \right], \qquad (1)$$

in which $p_{10}(f(x))$ is the 10th percentile of $f(x)$ and the factor κ is an empirically determined constant gain of 7. These factors (i.e., 10% shift and the factor κ) were empirically found to sufficiently capture the time window containing VF abductory and adductory gestures (within envelopes E^{H} and E^{L} overlap), while excluding data points corresponding to VF vibration.

VF data contained within the overlap (see Figure 3a) were then low-pass filtered with a Hamming window-based finite impulse response filter of order $n = 15$ with a cutoff frequency of 25 Hz to smooth the response for velocity estimation. From here, maximum adductory angular velocity was computed by finding the maximum declination rate located between the maximum glottic angle and 20% of the maximum glottic angle. This 20% limit was first adopted by Cooke et al. (1997) as a method of normalizing glottic trajectories across gestures. In particular, the authors demonstrated that termination of adduction was difficult to identify in certain cases, such as when vibration begins concurrently with the adduction process (Cooke et al., 1997). Although identifying the start of vibration is more straightforward when using HSV (e.g., via frame-by-frame analysis; Deliyski, 2010), we adopted the same 20% limit for our nonparametric analysis to account for potential abrupt VF closure (e.g., when vibratory onset precedes adduction termination). Figure 3b shows an example of the resulting glottic angle and maximum adduction velocity waveforms using nonparametric analysis.

Parametric analysis. Adduction trajectories and maximum adduction velocity were estimated from asymmetric sigmoidal fits derived from down-sampled glottic angle data (see Figure 4); the angle data were down-sampled in order to emulate lower frame rate scenarios. Only down-sampled data points within the adduction window remained as part of the adduction-sampled data. VF adduction trajectories were then estimated by fitting a four-parameter asymmetric sigmoid model onto the remaining angle data points (Equation 2), known as a Gompertz function (Tan, 2008).

$$f(t; a, b, c, d) = d + (a - d) \cdot 2^{-\mathrm{e}^{-b(t-c)}} \qquad (2)$$

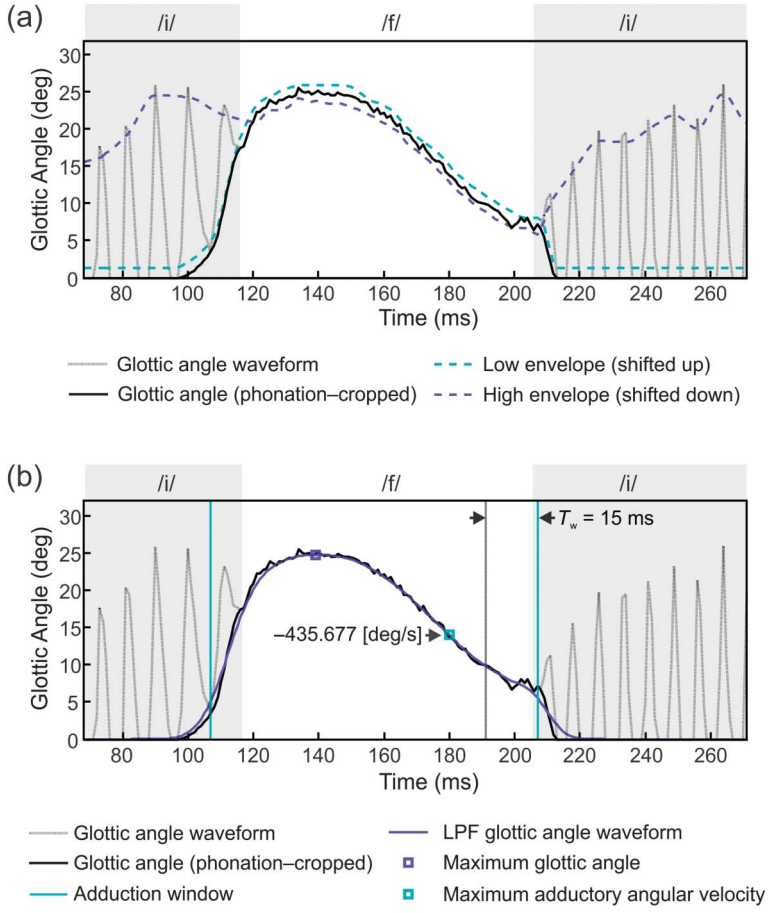
Maximum adduction angular velocity was then estimated by extracting the maximum declination rate of the sigmoid curve in Equation 2. The use of a Gompertz function was motivated by the need of a asymmetric sigmoidal curve with parameters flexible enough to (a) accommodate the location and extent of motion of the VF angle observed (i.e., maximum and minimum angle asymptotes), (b) implement an inflection point that is adjustable but limited to extent of motion so the function can be fitted over the asymmetric S-shape observed at adduction times, and (c) reduce the number of parameters needed to estimate the VF trajectory. The Gompertz model is a specific case of the generalized logistic function; however, the Gompertz function requires one less parameter to be described, which empirically proved to be numerically more stable for our purposes than the more general form. The sigmoidal fitting process is described in detail at its respective subsection below.

Frame rate simulation. Glottic angle trajectories obtained at 1,000 fps were down-sampled to emulate those that could be obtained at slower frame rate scenarios. Down-sampling was chosen in order to minimize potential measurement errors that may be introduced by tuning the angle extraction algorithm to process resampled video inputs. With respect to the latter point, changes in frame rate could affect internal or user-defined parameter selections; as a result, algorithmic performance may not be comparable across frame rate conditions when using resampled video inputs. Instead, emulating lower frame rates allows for the direct comparison of angles within a respective VCV instance across different frame rates.

All glottic angle waveforms considered usable during the angle extraction procedure were down-sampled progressively from 1,000 fps to 480, 240, 120, 60, and 30 fps. These values were empirically chosen to match the conventional endoscopic frame rate of 30 fps. The down-sampling phase was randomized to remove synchrony between down-sampled sets of the same /ifi/ instances. The 1,000-fps waveforms were considered a reference and were processed using both nonparametric and parametric techniques shown in Figures 3 and 4, respectively.

Sigmoidal fitting process. A Gompertz sigmoid function was implemented as a target model to characterize the adduction trajectories resulting from the parametric analysis. In cases where only a small number of kinematic samples are available to model the adduction trajectory, prior work (McKenna et al., 2016) suggests implementing fitting parameters or assuming asymptotic values outside the adduction time frame. As a result, five sets of reasonable assumptions were chosen for the target model under down-sampled conditions; these sets are denoted as *fitting strategies*. Each of these fitting strategies reflects a priori

Figure 3. Nonparametric glottic angle analysis, with (a) low/high envelope thresholding to isolate abduction and adduction trajectories and (b) filtering of the phonation-cropped glottic angle waveform, with maximum angular velocity calculated for the adductory gesture. The length of the filter (order $n = 15$) is shown by a filter window size (black arrows) of $T_w = 15$ ms. Shaded regions represent the duration of time corresponding to the vowel /i/ in the utterance /ifi/.



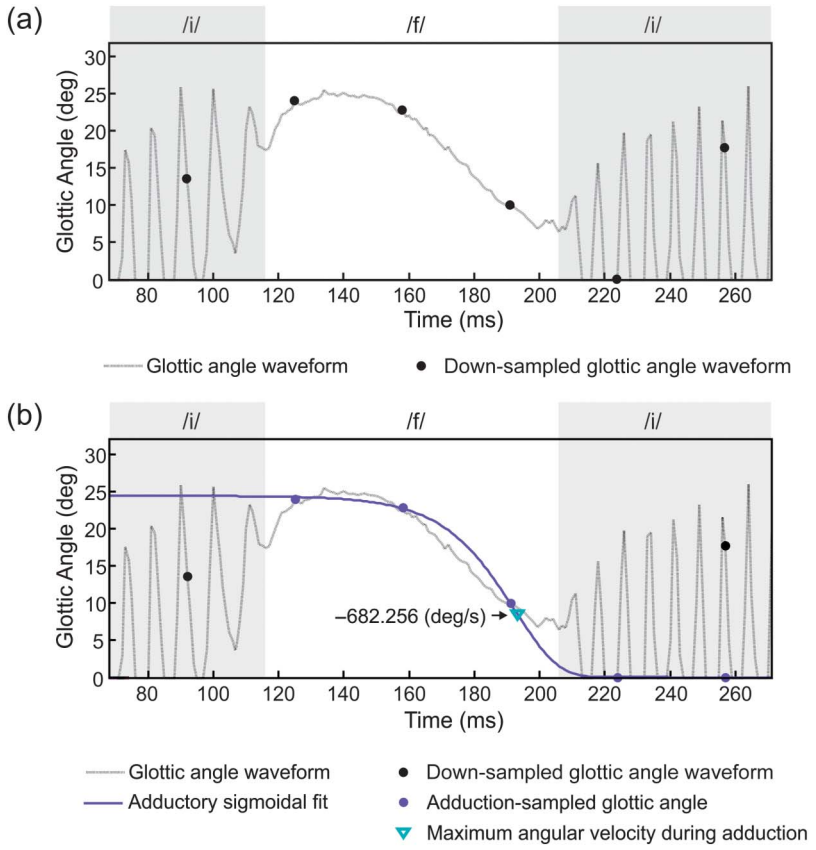
expectations for the adduction trajectory behavior. For instance, a fitting strategy could fix the offset parameter of the target model to zero to assume a minimum glottic angle of zero. Conversely, this assumption could be encoded by extending the adduction trajectory with additional zero-valued data points, such that the sigmoidal fit favors zero. Similar assumptions may be proposed for maximum glottic angles in order to restrict the fitting solution at the initiation of adduction.

Different fitting strategies were implemented to reflect assumptions made regarding the sigmoid curve during adduction; each strategy employs the same sigmoid model, but with different restrictions applied. Different fitting strategies were implemented to reflect assumptions made regarding the sigmoid curve during adduction; each strategy employs the same sigmoid model, but with different restrictions applied. These fitting strategies were implemented to reflect a variety of methods that may be used to estimate the timing and final motions of the VFs during the adduction process. The motivation for these strategies was derived from previous observations performed by Cooke et al. (1997); in particular, this study described the final stages of the adductory gesture as difficult to observe because (a) VF trajectory transitions from adduction to phonation are a continuous process, wherein the timing for final VF posture could be occluded by supraglottic compression or preceded by the very first onset of phonation cycles, and (b) the adductory gesture may end in a complete or incomplete glottal closure. As such, five possible fitting strategies (S1–S5) were evaluated (see Figure 5) to consider these observations in the parametric estimation of the adductory trajectory:

S1. The offset and amplitude parameters of the fitted trajectory (Equation 2) are set to zero and to the value of the maximum detected glottic angle, respectively. This strategy assumes that adduction always ends with full glottic closure and that the maximum asymptotic value of the sigmoid fit is equal to the maximum observed glottic angle.

S2. Offset and amplitude parameters of the fitted trajectory mimic that of S1. Additionally, the portion of the

Figure 4. Parametric glottic angle analysis, with (a) down-sampled glottic angle waveform estimated at 1,000 frames per second to simulate lower frame rates and (b) fitting an asymmetric sigmoid model to simulated data points within the adduction window, with maximum angular velocity calculated for the adductory sequence. Shaded regions represent the duration of time corresponding to the vowel /i/ in the utterance /ifi/.



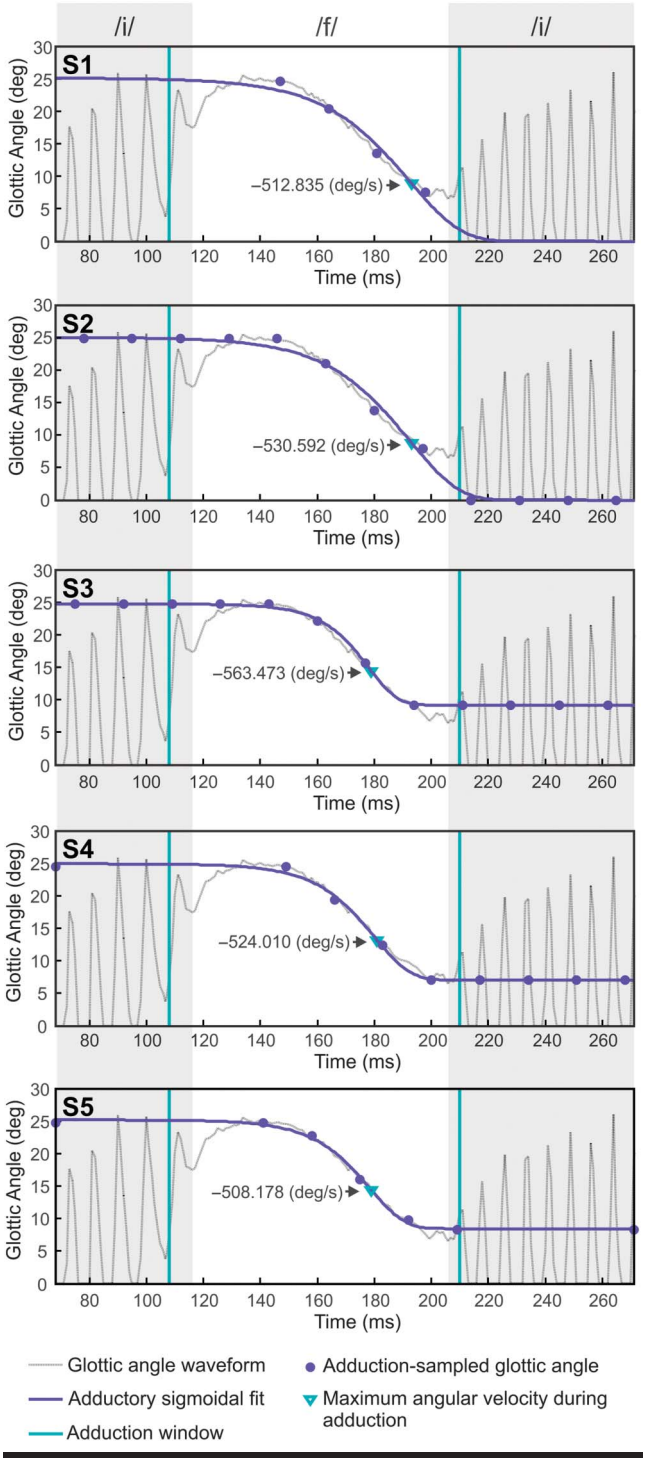
VF trajectory prior to the initiation of adduction is appended with the maximum angle value, while the portion of the VF trajectory considered after the end of adduction is appended with a zero value. This strategy assumes the same restrictions as S1; however, a stronger weighting is applied to flatten the tails of the sigmoid curve outside the adduction window.

S3. The offset and amplitude parameters of the fitted trajectory are set to the value of the minimum detected glottic angle and to the value of the maximum detected glottic angle, respectively. The glottic angle trajectory is appended using the value of the maximum angle prior to adduction and of the minimum angle following adduction. This strategy assumes similar restrictions as S2, except that there is no presumption that the adduction window ends with full glottic closure.

S4. The offset parameter of the fitted trajectory is forced to the value of the minimum detected glottic angle found within the adduction window; however, the amplitude parameter is estimated via the sigmoidal fitting process. The fitted trajectory is appended with a single data point that is valued at the maximum glottic angle prior to the adduction window. The glottic angle trajectory is then extended following the end of adduction using minimum angle data points. The assumption in this strategy is different from previous strategies with respect to the initiation of adduction, in that the maximum observed glottic angle is not the value at which the sigmoid curve begins. The fitting process must then find the amplitude parameter using the maximum glottic angle as an initial estimate, while the appended data point applies an additional weight to keep the optimal amplitude parameter within a similar magnitude to the initial estimate.

S5. All sigmoid parameters are estimated during the sigmoidal fitting process. The glottic angle trajectory is extended on either end of the adduction window using only one data point: a maximum angle data point at the start of adduction and a minimum glottic data point at the end of adduction. This strategy applies similar techniques as S4 for adduction initiation and termination; yet, S5 also considers that the minimum observed glottic angle during the adduction window does not correspond to the exact value that the sigmoid asymptotically approaches at the end of the window.

Figure 5. Example of fitting strategies S1–S5 (denoted in the upper-left corner). All down-sampled waveforms emulate video recording at 60 frames per second. Blue circles correspond to the altered glottic angle waveform after taking into account the replacement, addition, or removal of down-sampled data according to fitting strategy. Shaded regions represent the duration of time corresponding to the vowel /i/ in the utterance /ifi/.



Statistical Analysis

Two repeated-measures analysis of variance tests were performed using restricted maximum likelihood estimation to analyze the effects of fitting strategy and frame rate on MAV and RMS errors of the adduction trajectories. Prior to analysis, the magnitude of MAV and RMS errors were log transformed to meet the assumption of normality. For each analysis of variance model, participant was set as a random factor with simulated sampling rate (six levels: 30, 60, 120, 240, 480, and 1,000 fps), strategy (five levels: S1, S2, S3, S4, and S5), condition (six levels: SR, RR, FR, MIL, MOD, and MAX), and all interactions as fixed effects. An α level of .05 was used as a cutoff criterion for significance in each analysis. Effect sizes for the factors were calculated using a squared partial curvilinear correlation (η_p^2).

Results

Effects of Condition, Fitting Strategy, and Simulated Frame Rate on Adduction Trajectory Errors

Table 1 displays the model summaries constructed for RMS and MAV errors. Condition, simulated frame rate, and strategy each had a significant effect on the log-transformed MAV and RMS errors ($p < .001$ for all). Additionally, interactions of Condition × Simulated Frame Rate, Condition × Strategy, and Simulated Frame Rate × Strategy were significant in both models. Of note, the interaction of Condition × Simulated Frame Rate × Strategy was not significant for either variable.

Simulated frame rate was observed to have a large effect size ($\eta_p^2 = .34$ for RMS error, $\eta_p^2 = .65$ for MAV error) on the resulting adduction trajectory errors. The interaction of Simulated Frame Rate × Strategy had a medium effect size on RMS error ($\eta_p^2 = .12$) and a large effect size on MAV error ($\eta_p^2 = .35$). Strategy was observed to produce medium effects on the outcome measures ($\eta_p^2 = .17$ for RMS error, $\eta_p^2 = .11$ for MAV error). Although condition was significant in both models, only a small amount of the variance in either RMS or MAV error was attributable to condition or its interactions with simulated frame rate and strategy.

Adduction Trajectory Errors by Fitting Strategy

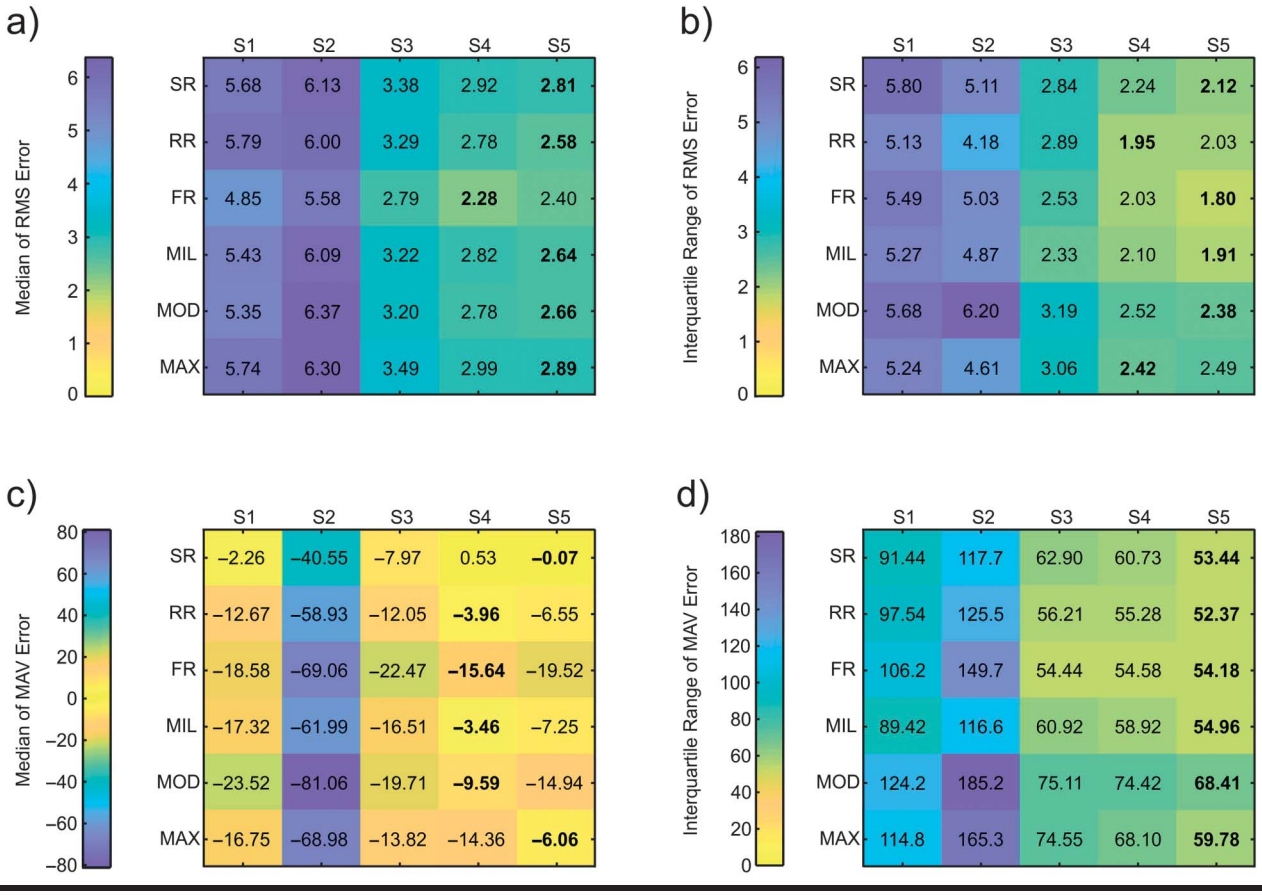
Figure 6 displays median and interquartile range (IQR) values for RMS and MAV errors across speech condition and fitting strategy. Error values were calculated by comparing parametric and nonparametric results at a down-sampled frame rate of 120 fps; a frame rate of 120 fps was selected for this analysis since it was the median simulated frame rate of those examined (i.e., 30, 60, 120, 240, and 480 fps). Median RMS error was lowest in strategies without a zero-level restriction for the offset parameter (i.e., S3, S4, and S5), regardless of speech condition (see Figure 6a). Conversely, median MAV error changed

Table 1. Results of analysis of variance tests on root-mean-square (RMS) and maximum angular velocity (MAV) errors between parametric and nonparametric analyses.

Model	Effect	*df*	η_p^2	*F*	*p*
$\log_{10}$|RMS error|	Participant	24	.23	587.6	< .001
	Condition	5	.01	53.1	< .001
	Simulated frame rate	5	**.34**	5004.0	< .001
	Strategy	4	**.17**	2392.0	< .001
	Condition × Simulated Frame Rate	25	.00	4.5	< .001
	Condition × Strategy	20	.00	2.6	< .001
	Simulated Frame Rate × Strategy	20	**.12**	336.0	< .001
	Condition × Simulated Frame Rate × Strategy	100	—	0.4	1.00
$\log_{10}$|MAV error|	Participant	24	.27	148.3	< .001
	Condition	5	.04	73.9	< .001
	Simulated frame rate	5	**.65**	3733.0	< .001
	Strategy	4	**.11**	308.8	< .001
	Condition × Simulated Frame Rate	25	.01	3.8	< .001
	Condition × Strategy	20	.00	1.7	.028
	Simulated Frame Rate × Strategy	20	**.35**	262.7	< .001
	Condition × Simulated Frame Rate × Strategy	100	—	0.8	.918

Note. Effect sizes are only reported for statistically significant factors. Bolded effect sizes are those discussed further in the article. Em dashes signify the lack of reporting.

Figure 6. Error values of parametric and nonparametric analyses by condition (vertical axis) and strategy (horizontal axis) at a frame rate of 120 fps, with (a) median root-mean-square (RMS) error, (b) interquartile range of RMS error, (c) median of maximum adduction velocity (MAV) error, and (d) interquartile range of MAV error. Error values equal to zero (yellow) represent complete correspondence between the parametric and nonparametric analyses. Bolded values indicate the smallest errors per condition. FR = fast rate; MAX = maximum; MIL = mild; MOD = moderate; RR = regular rate; SR = slow rate.



as a function of sigmoidal fitting strategy and speech condition. Yet, considerable differences were not observed in median MAV errors across condition, which is likely a result of high variability in error (see Figure 6d). The IQRs of MAV errors were, however, found to be considerably reduced in strategies without a zero-level restriction (i.e., S3–S5) when compared to those with such a restriction imposed (i.e., S1–S2).

Figure 7 shows the median and IQR values for RMS and MAV errors across frame rate and fitting strategy. Error values were calculated by comparing parametric results for each frame rate with nonparametric results at 1,000 fps. When examining trajectory errors across frame rate, median and IQR values of RMS error were lower in strategies that did not impose a zero-level restriction (i.e., S3–S5) for frame rates above 30 fps (i.e., 60–1,000 fps; see Figures 7a and 7b); however, the opposite trend was observed for RMS errors at 30 fps. Additionally, MAV errors were, on average, lower in magnitude for these strategies at 60 fps or higher. Of note, trajectory errors increase substantially when parametric analysis is conducted using a down-sampled rate of 30 fps.

Adduction Trajectory Errors Between Conventional and High-Speed Frame Rates

Figures 8 and 9 compare results of parametric and nonparametric analyses between frame rates of 30 and 1,000 fps. These two frame rates were chosen to compare since 30 fps has been considered the standard speed for videoendoscopy systems (Britton et al., 2014, 2012; Dailey et al., 2005; McKenna et al., 2016; Stepp et al., 2010), whereas 1,000 fps was the highest sampling rate employed in the current study.

When comparing parametric and nonparametric analyses using 1,000 fps (see Figure 8a, right boxplots at each strategy), median RMS errors were larger when the sigmoid fit was forced to zero at the termination of adduction (S1 and S2) rather than leaving the respective parameter to the fitting process (S5). On average, median MAV error was relatively stable across strategy for 1,000-fps

Figure 7. Error values of parametric and nonparametric analyses by frame rate (vertical axis) and strategy (horizontal axis), with (a) median root-mean-square (RMS) error, (b) interquartile range of RMS error, (c) median of maximum adduction velocity (MAV) error, and (d) interquartile range of MAV error. Error values equal to zero (yellow) represent complete correspondence between the parametric and nonparametric analyses. Bolded values indicate the smallest errors per frame rate. fps = frames per second.

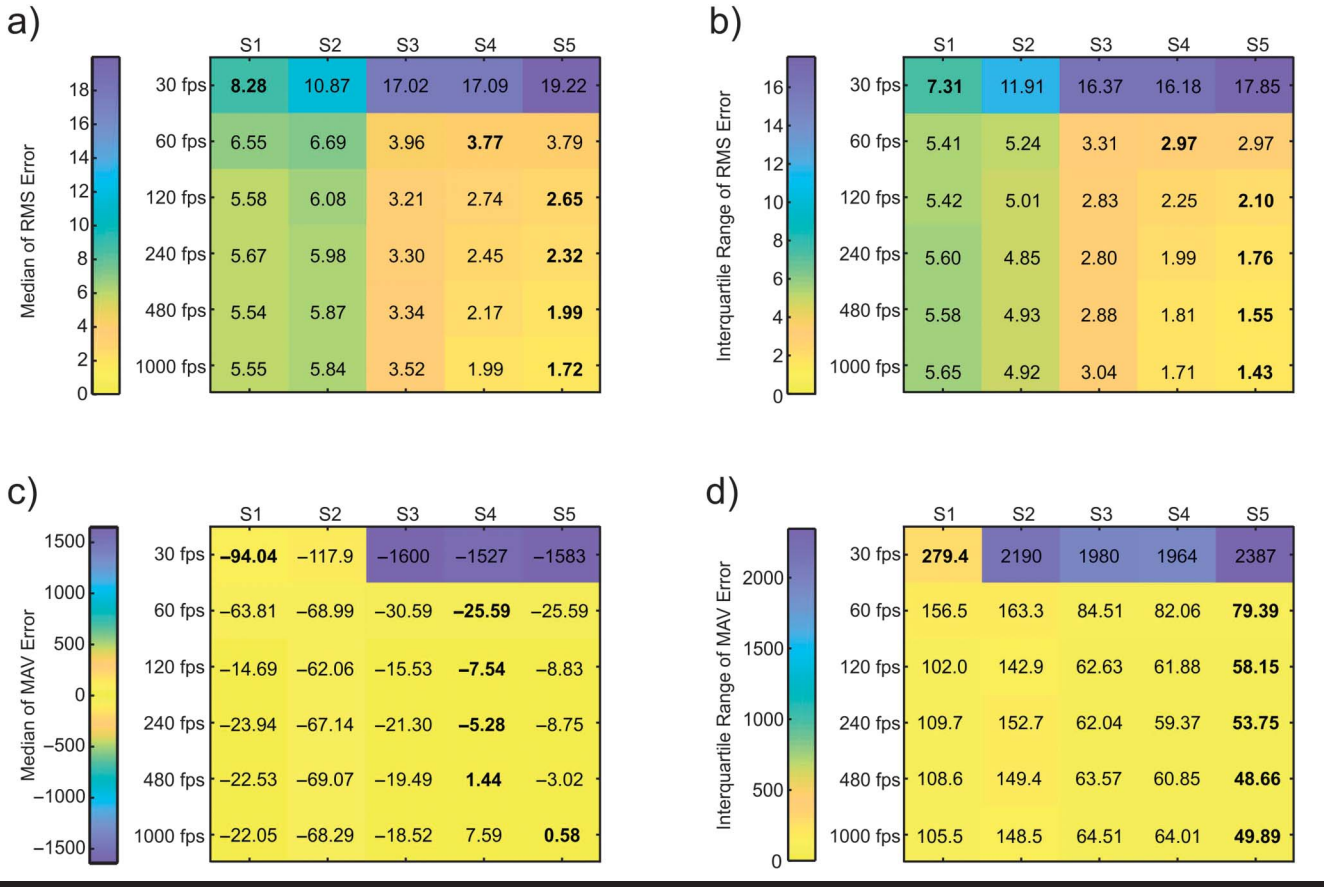
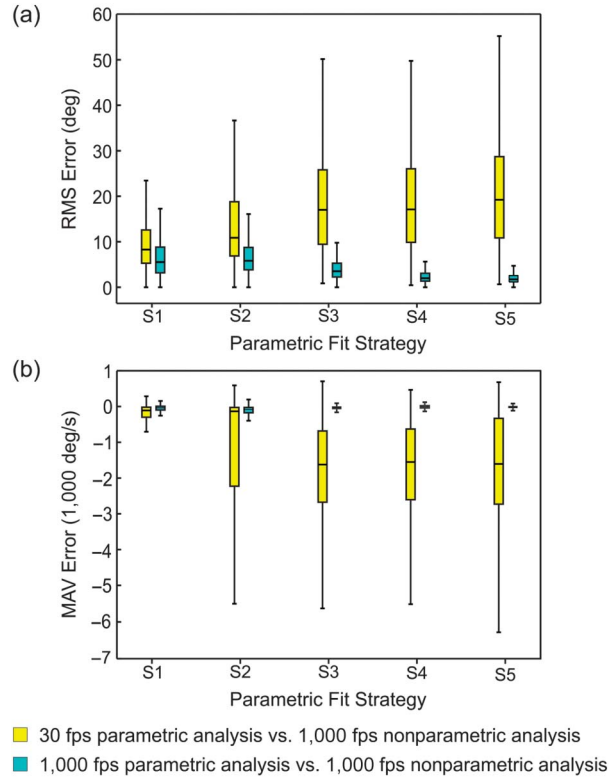


Figure 8. Boxplot comparison of adduction trajectory errors across strategy. Two analyses are compared: 30 frames per second (fps) parametric versus 1,000 fps nonparametric and 1,000 fps parametric versus 1,000 fps nonparametric. Root-mean-square (RMS) errors are shown in (a), and maximum adduction velocity (MAV) errors are shown in (b).



analyses (middle boxplots at each strategy in Figure 8b). Such trends are not observed when comparing the 30-fps parametric results with the 1,000-fps nonparametric results (left boxplots). In particular, the median and IQR of RMS and MAV error each increase across strategy, with zero-level restriction strategies (i.e., S1 and S2) producing the smallest errors.

Compared with S5, S1 showed a relatively small change in median error as simulated frame rate increased (left boxplots at each frame rate in Figure 9), changing from median RMS error of 8.28° to 5.55° (see Figure 9a) and respective median MAV errors from −94.04 to −22.05 deg/s (see Figure 9c). In contrast, S5 showed a larger range of improvement in median RMS error, changing from 19.22° to 1.72° (see Figure 9b), and median MAV errors from −1,583 to 0.58 deg/s (see Figure 9d).

Sigmoidal fitting often failed when extended data point strategies were not used to fit 30-fps data (i.e., S5 against S1–S4); in these scenarios, sigmoidal trajectory estimates were incalculable approximately 33% of the time. Yet, despite the large failure rate, fitting procedures that append extra data points to the fitted adduction trajectories may still be too unstable for implementation. Specifically, the IQRs of RMS and MAV errors were larger for S2–S5 despite successful sigmoid fitting (see Figures 9b and 9d). Of note, fit failures occurred less often when the frame rate was increased above 30 fps (less than 5% as shown in Figures 9e and 9f).

At lower frame rates, all sigmoid fit strategies overestimated maximum angle velocities (negative median MAV errors observed in Figures 9c and 9d). However, RMS and MAV errors were reduced as the frame rate increased, although this reduction was not linear. It can be observed that an error inflection point occurs at 120 fps, in which median RMS and MAV errors were relatively stable for both fitting strategies (S1 and S5) at frame rates beyond this inflection point.

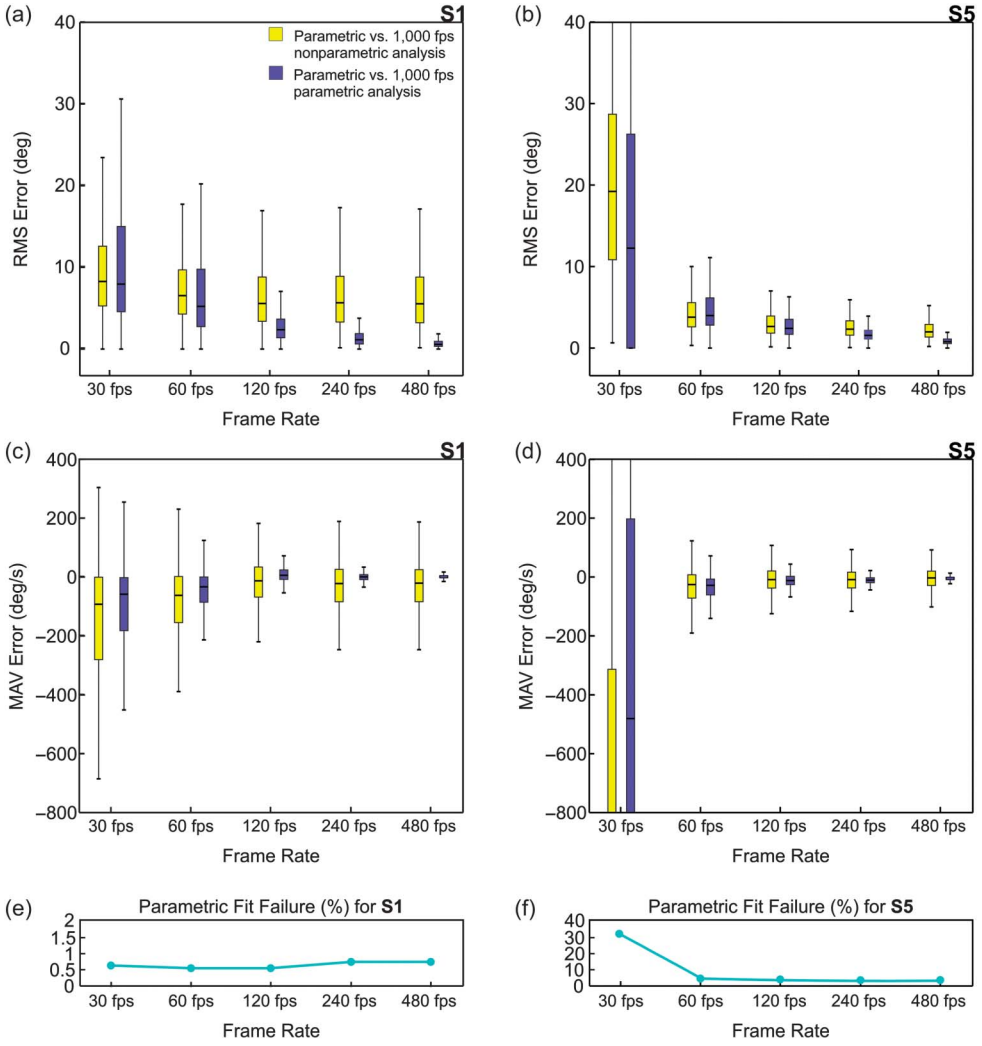
Discussion

Modeling the Behavior of Adduction Trajectories as Sigmoidal Functions

The present findings are consistent with previous work by Iwahashi et al. (2016), in which the closing gesture can be seen decelerating prior to VF impact at phonation onset. Similarly, in the present data, monotonically decreasing angle trajectories during adduction were consistently observed. Therefore, an asymmetric sigmoid model is a plausible representation of the average glottic angle behavior during adduction (i.e., prior to phonation) as long as (a) the VCV instance presents a monotonically decreasing curve and (b) the number of samples and fitting strategy are reasonably chosen. In general, however, fits to adduction trajectories overestimate MAVs when the frame rate is decreased; as such, fitting failures are expected to increase substantially below 60 fps. It follows that asymmetric sigmoid models may be less useful in describing VF adduction trajectories obtained from conventional videoendoscopies obtained at 30 fps, despite the fact that it is at these lower frame rates that the sparse data require such fitting techniques the most. We observed that a frame rate of at least 120 fps was required to increase the predicting power of the model.

Prior work in this area has already determined that VF assessment via HSV is sensitive to frame rate. For instance, Popolo (2017) concluded that clinical rates based on mucosal wave features have the potential to create discrepancies in clinical assessments under inadequate frame rates; a minimum of 4,000 fps was considered necessary to minimize feature degradation. However, this work focuses on observing slower laryngeal motions unrelated to mucosal wave features. Instead, abduction and adduction kinematics originate by varying laryngeal muscular activation to move the laryngeal structures. Adduction trajectories for the VCV utterance /ifi/, described here, occur over approximately 150 ms (McKenna et al., 2016). Since the nature of these motions is driven to a greater degree by muscle activation rather than by aerodynamics, slow motion trajectories may be captured at 1,000 fps. As a result, the adduction model-based estimations remain valid even

Figure 9. Comparison of adduction trajectory errors for parametric analysis performed at low frame rates (i.e., 30, 60, 120, 240, and 480 frames per second [fps]) versus 1,000-fps nonparametric analysis (left boxplots at each frame rate) and 1,000-fps parametric analysis (right boxplots at each frame rate). Root-mean-square (RMS) errors are shown in (a) and (b), maximum adduction velocity (MAV) errors are shown in (c) and (d), and parametric fit failure percentages are shown in (e) and (f). Errors and fit failures for S1 are shown in (a), (c), and (e), while those for S5 are shown in (b), (d), and (f).



after down-sampling from 1,000 fps to simulated scenarios of 480 and 240 fps.

Effects of Fitting Strategy on Adduction Model-Based Estimation

Selecting the most appropriate fitting strategy substantially impacts the associated fitting error. The most valid fitting strategy for the lowest frame rate (i.e., 30 fps) is not the best strategy for frame rates of 60 fps or higher: As soon as more data points are available for the sigmoid fit, the resulting median error considerably decreases when switching from S1—which resulted in the lowest error when a frame rate of 30 fps was used—to another strategy (S2–S5). A particular example of this can be examined by comparing the fitted curves in Figure 5 with the reference curve in Figure 3b. Specifically, the adduction curve fits the extracted glottic angle waveform better when using S5 rather than S1 at 60 fps; however, reducing the sampling rate to 30 fps negatively impacts S5, such that the fit process fails due to a lack of data points for the regression. In contrast, S1 performs poorly against the reference, but it is still calculable with the few points available (see Figure 4b).

This variation in strategy performance may be a result of S5 estimating two additional parameters that S1 considers known prior to the sigmoidal fitting process. Specifically, the number of unknown parameters in the model dictates the minimum data points needed to compute a nonlinear regression. Given that less data points are available for the sigmoidal fitting process at lower frame rates,

parametric fit failures (see Figure 9f) below 60 fps are substantially more likely to occur when compared to higher frame rates.

Additionally, two distinguishable performance groups were observed according to how the offset parameter was selected. In particular, strategies that forced the minimum adductory angle of the fitted trajectory to zero produced different results when compared to strategies that forced this parameter to the value of the minimum detected glottic angle found within the adduction window. Therefore, the former were referred to as *zero-level* strategies (S1 and S2), whereas the latter corresponded to *minimum-angle* strategies (S3, S4, and S5). The predictive power of the minimum-angle strategies was systematically better than that of the zero-angle strategies at frame rates of 60 fps or higher. This can be observed in Figures 7a and 7c, wherein resulting median errors are smaller for S3–S5 when compared to S1 and S2. These findings are in contrast to work by Iwahashi et al. (2016), in which representative angle trajectories for throat clearing and steady-phonation onset continued to decrease within the 20%–0% range, ultimately reaching a zero angle at the end of the motion. Instead, the observed glottal angles in this work failed to reach zero in the vast majority of instances. This is likely a result of different study tasks: The current study examined VCV utterances, whereas Iwahashi et al. examined throat clearing and steady phonation. Regardless, our findings suggest that zero-level strategies are less suitable to represent the observed glottic angles. As such, fitting strategies that estimate the offset parameter (i.e., minimum-angle strategies) should be implemented when possible to analyze glottic angles using a sigmoid model. While the zero-level strategy, S1, is recommended for analyzing videoendoscopic data recorded at 30 fps, this fitting strategy should be avoided when analyzing data at higher frame rates due to considerable estimation errors attributed to the invalid zero-angle assumption.

Minimum angle discrepancies may be a result of differences in the nature of the task performed in the current investigation. As previously mentioned, the majority of participants in this study failed to reach complete glottal closure between the end of adduction and the start of the phonation during VCV productions. Iwahashi et al. (2016) extracted glottic angles starting at a rest abductory position and throughout sustained phonation; it is likely that participants performed an initial inhalation or other preparatory gestures prior to phonation. Conversely, individuals in this study continuously exhaled air to produce the consonant /f/ in between two vowels /i/ within the instructed VCV utterances. As a result, there were no airflow-based pauses. The continuous speech necessary to complete the tasks in the current study may then have failed to elicit minimum adductory angles at zero.

Clinical Significance

These recommendations can be extended to clinical assessments using videoendoscopy, wherein implementing HSV poses numerous challenges (see Deliyski et al., 2008, for a comprehensive review). These challenges include practical and technical issues, such as storing large amounts of data that fail to span multiple phonatory productions (i.e., short sample durations). However, our results suggest that, when conventional frame rates of 30 fps are used to evaluate VF closing velocities, fitting strategy S1 should be used. Previous studies using VF adductory kinematics that were acquired using conventional speed videoendoscopy systems at a frame rate of 30 fps (Britton et al., 2014, 2012; Dailey et al., 2005; McKenna et al., 2016; Stepp et al., 2010) have not generally reported fitting strategy. It is possible that the results of these studies may have been impacted by fitting errors. Future work should examine the study questions of each with respect to the possible influences of frame rate and fitting strategy.

In the event that higher frame rates are clinically available, our findings suggest that frame rates of 1,000 fps or higher are not required to accurately capture kinematic behavior; indeed, frame rates as low as 120 fps can be implemented, with resulting videoendoscopic data fit to an asymmetric sigmoidal model using S5. This suggests that (a) longer videos can be captured to include multiple phonatory productions or (b) shorter videos can be captured to account for clinical storage considerations. Although current challenges for implementing HSV in the clinic span practicality, technicality, methodology, and clinical applicability, the findings in this study improve the prospects of using HSV for the clinical assessment of VF kinematic behavior.

Limitations and Future Directions

All of the presented interpretations for the model-based estimation of adduction trajectories are limited to typical, young healthy speakers. Further conclusions cannot be formulated for VF trajectories that do not present a clear, monotonically decreasing curve at the initiation of adduction. Potential adduction irregularities resulting from partial VF paralysis, dysphonia, tremor, or other considerable VF trajectory perturbations (not necessarily pathological) may also invalidate this assumption of a monotonic curve. Previous model-based VF kinematic analyses have assumed a monotonically decreasing trajectory for tasks such as cough (Britton et al., 2012), breath-interrupted phonations (e.g., /i/-sniff; Dailey et al., 2005; Stepp et al., 2010), or VCV segments (e.g., /ifi/; McKenna et al., 2016), with the aim of estimating angular adduction velocity with fewer data samples (e.g., 30 fps); however, other models should be investigated to estimate trajectory behaviors in more complex voice productions. Although implementing a nonparametric method requires a higher frame rate, it is likely to be more adequate in such investigations since a priori expectations about VF trajectory are avoided. Furthermore, the current study examined high-speed videos recorded at 1,000 fps; however, it is possible that our findings do not translate to videos recorded at higher frame rates. Finally, angle extraction errors due to glottic angle estimation can propagate to the sigmoid model analysis:

A noisy and inaccurate glottic angle waveform may be generated in cases where video processing struggles due to poor input quality. This waveform only serves to add uncertainty in analysis results. Therefore, improvements in camera resolution and lighting, in addition to advances in endoscope technology, may improve these deficits in future studies.

Conclusions

In this study, VF angle trajectories during adduction were examined in young adults with healthy voices. We showed that an asymmetric sigmoidal model may be a useful way to describe the kinematic behavior during VF adduction. To summarize, our findings suggest that adduction trajectory estimates are similar to the sigmoidal fit if a frame rate of at least 120 fps is implemented in conjunction with fitting strategy S5. We recommend this method as the minimal settings needed to describe VF angle trajectories of VCV utterances with reasonable estimation bias and error variability. However, if the conventional sampling rate of 30 fps is chosen to examine kinematic behavior, then fitting strategy S1 should be used to evaluate VF angle trajectories. Although implementing HSV as a clinical technique poses various challenges, the findings in this study (a) demonstrate a viable approach for assessing VF kinematic behavior at frame rates as low as 120 fps and (b) provide the appropriate fitting strategies for using conventional 30 fps to evaluate closing velocities in a clinical setting.

Acknowledgments

This work was supported by Grants DC015570 (awarded to Cara E Stepp) and DC013017 (awarded to Christopher A Moore) from the National Institute on Deafness and Other Communication Disorders and National Science Foundation Graduate Research Fellowship Grant 1247312 (awarded to Jennifer M. Vojtech). The authors would like to thank Jacob Noordzij, Jr., Jaime Kim, and Lin Zhang for assistance with data processing and Adrianna Shembel for assistance with data acquisition.

References

Aghdam, M. A., Ogawa, M., Iwahashi, T., Hosokawa, K., Kato, C., & Inohara, H. (2017). A comparison of visual recognition of the laryngopharyngeal structures between high and standard frame rate videos of the fiberoptic endoscopic evaluation of swallowing. *Dysphagia, 32*(5), 617–625.

Braunschweig, T., Flaschka, J., Schelhorn-Neise, P., & Döllinger, M. (2008). High-speed video analysis of the phonation onset, with an application to the diagnosis of functional dysphonias. *Medical Engineering & Physics, 30*(1), 59–66.

Britton, D., Benditt, J. O., Merati, A. L., Miller, R. M., Stepp, C. E., Boitano, L., . . . Yorkston, K. M. (2014). Associations between laryngeal and cough dysfunction in motor neuron disease with bulbar involvement. *Dysphagia, 29*(6), 637–646.

Britton, D., Yorkston, K. M., Eadie, T., Stepp, C. E., Ciol, M. A., Baylor, C., & Merati, A. L. (2012). Endoscopic assessment of vocal fold movements during cough. *Annals of Otology, Rhinology & Laryngology, 121*(1), 21–27.

Cooke, A., Ludlow, C. L., Hallett, N., & Scott Selbie, W. (1997). Characteristics of vocal fold adduction related to voice onset. *Journal of Voice, 11*(1), 12–22.

Dailey, S. H., Kobler, J. B., Hillman, R. E., Tangrom, K., Thananart, E., Mauri, M., & Zeitels, S. M. (2005). Endoscopic measurement of vocal fold movement during adduction and abduction. *Laryngoscope, 115*(1), 178–183.

Deliyski, D. D. (2010). Laryngeal high-speed videoendoscopy. In K. A. Kendall & R. J. Leonard (Eds.), *Laryngeal evaluation* (pp. 245–259). New York, NY: Laryngeal Evaluation.

Deliyski, D. D., Petrushev, P. P., Bonilha, H. S., Gerlach, T. T., Martin-Harris, B., & Hillman, R. E. (2008). Clinical implementation of laryngeal high-speed videoendoscopy: Challenges and evolution. *Folia Phoniatrica et Logopaedica, 60*(1), 33–44.

Deliyski, D. D., Powell, M. E., Zacharias, S. R., Gerlach, T. T., & de Alarcon, A. (2015). Experimental investigation on minimum frame rate requirements of high-speed videoendoscopy for clinical voice assessment. *Biomedical Signal Processing and Control, 17,* 21–28.

Döllinger, M., Dubrovskiy, D., & Patel, R. R. (2012). Spatiotemporal analysis of vocal fold vibrations between children and adults. *Laryngoscope, 122*(11), 2511–2518.

Freeman, E., Woo, P., Saxman, J. H., & Murry, T. (2012). A comparison of sung and spoken phonation onset gestures using high-speed digital imaging. *Journal of Voice, 26*(2), 226–238.

Guzman, M., Laukkanen, A. M., Traser, L., Geneid, A., Richter, B., Muñoz, D., & Echternach, M. (2017). The influence of water resistance therapy on vocal fold vibration: A high-speed digital imaging study. *Logopedics Phoniatrics Vocology, 42*(3), 99–107.

Hetrich, I., & Ackermann, H. (1995). Coarticulation in slow speech: Durational and spectral analysis. *Language and Speech, 38*(Pt. 2, 159–187.

Ikuma, T., Kunduk, M., Fink, D., & McWhorter, A. J. (2016). A spatiotemporal approach to the objective analysis of initiation and termination of vocal-fold oscillation with high-speed videoendoscopy. *Journal of Voice, 30*(6), 756.e21–756.e300.

Ishii, I., Takemoto, S., Takaki, T., Takamoto, M., Imon, K., & Hirakawa, K. (2011). *Real-time laryngoscopic measurements of vocal-fold vibrations*. Paper presented at the 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (pp. 6623–6626). https://doi.org/10.1109/IEMBS.2011.6091633

Iwahashi, T., Ogawa, M., Hosokawa, K., Kato, C., & Inohara, H. (2016). A detailed motion analysis of the angular velocity between the vocal folds during throat clearing using high-speed digital imaging. *Journal of Voice, 30*(6), 770.e1–770.e8.

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

Kunduk, M., Döllinger, M., McWhorter, A. J., & Lohscheller, J. (2010). Assessment of the variability of vocal fold dynamics within and between recordings with high-speed imaging and by phonovibrogram. *Laryngoscope, 120*(5), 981–987.

Kunduk, M., Vansant, M. B., Ikuma, T., & McWhorter, A. (2017). The effects of the menstrual cycle on vibratory characteristics of the vocal folds investigated with high-speed digital imaging. *Journal of Voice, 31*(2), 182–187.

Kunduk, M., Yan, Y., McWhorter, A. J., & Bless, D. (2006). Investigation of voice initiation and voice offset characteristics with high-speed digital imaging. *Logopedics Phoniatrics Vocology, 31*(3), 139–144.

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

Lohscheller, J., Toy, H., Rosanowski, F., Eysholdt, U., & Döllinger, M. (2007). Clinically evaluated procedure for the reconstruction of vocal fold vibrations from endoscopic digital high-speed videos. *Medical Image Analysis, 11*(4), 400–413.

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

Mehta, D. D., Deliyski, D. D., Quatieri, T. F., & Hillman, R. E. (2011). Automated measurement of vocal fold vibratory asymmetry from high-speed videoendoscopy recordings. *Journal of Speech, Language, and Hearing Research, 54*(1), 47–54.

Mehta, D. D., & Hillman, R. E. (2012). Current role of stroboscopy in laryngeal imaging. *Current Opinion in Otolaryngology & Head and Neck Surgery, 20*(6), 429–436.

Ostry, D. J., & Munhall, K. G. (1985). Control of rate and duration of speech movements. *The Journal of the Acoustical Society of America, 77*(2), 640–648.

Patel, R. R., Dubrovskiy, D., & Döllinger, M. (2014). Characterizing vibratory kinematics in children and adults with high-speed digital imaging. *Journal of Speech, Language, and Hearing Research, 57*(2), S674–S686.

Patel, R. R., Forrest, K., & Hedges, D. (2017). Relationship between acoustic voice onset and offset and selected instances of oscillatory onset and offset in young healthy men and women. *Journal of Voice, 31*(3), 389.e9–389.e17.

Patel, R. R., Unnikrishnan, H., & Donohue, K. D. (2016). Effects of vocal fold nodules on glottal cycle measurements derived from high-speed videoendoscopy in children. *PLOS ONE, 11*(4), e0154586.

Patel, R. R., Walker, R., & Sivasankar, P. M. (2016). Spatiotemporal quantification of vocal fold vibration after exposure to superficial laryngeal dehydration: A preliminary study. *Journal of Voice, 30*(4), 427–433.

Popolo, P. S. (2017). Investigation of flexible high-speed video nasolaryngoscopy. *Journal of Voice, 33*(5), 529–537.

Stepp, C. E., Hillman, R. E., & Heaton, J. T. (2010). A virtual trajectory model predicts differences in vocal fold kinematics in individuals with vocal hyperfunction. *The Journal of the Acoustical Society of America, 127*(5), 3166–3176.

Tan, C. Y. (2008). *Symmetric and asymmetric sigmoid curves: A close look at their statistical, numerical and mathematical properties*. Paper presented at the Non-Clinical Statistics Conference 2008, Leuven, Belgium Retrieved from http://www.ncs-conference.org/2008/slides/24/2/CT.pdf

Watanabe, T., Kaneko, K., Sakaguchi, K., & Takahashi, H. (2016). Vocal-fold vibration of patients with Reinke's edema observed using high-speed digital imaging. *Auris Nasus Larynx, 43*(6), 654–657.

Woo, P. (2017). High-speed imaging of vocal fold vibration onset delay: Normal versus abnormal. *Journal of Voice, 31*(3), 307–312.

Yamauchi, A., Yokonishi, H., Imagawa, H., Sakakibara, K., Nito, T., Tayama, N., & Yamasoba, T. (2016). Quantification of vocal fold vibration in various laryngeal disorders using high-speed digital imaging. *Journal of Voice, 30*(2), 205–214.

Zacharias, S. R. C., Deliyski, D. D., & Gerlach, T. T. (2018). Utility of laryngeal high-speed videoendoscopy in clinical voice assessment. *Journal of Voice, 32*(2), 216–220.

Appendix (p. 1 of 8)

VF Estimation Algorithm

The full details of the semiautomated VF estimation algorithms are detailed here.

Event Detector

The event detector leverages the auxiliary (i.e., microphone or accelerometer) signal to identify the location of each VCV instance within a recording. In particular, time points corresponding to the following are extracted: voice onset and offset for each of the vowels in a VCV instance, midpoint of each of the vowels, and segments in time where the vocal folds may be abducted. These segments of potential vocal fold abduction are then implemented in subsequent processing steps to define a frame of reference that may be used to track the glottic area during the transition into and out of the voiceless consonant.

The event detector takes an auxiliary signal as input rather than the HSV recording in order to minimize the computational cost needed to import and analyze each frame of the HSV recording. For the same reason, onset and offset detection is not necessarily accurate since precise detection of onset and offset timings is not required; specifically, we are interested in the glottic angle waveform across the transition into and out of the voiceless consonant. Therefore, event location errors on the order of one to five cycles do not significantly impact the final glottic angle results.

Equations 3–5 show the computations used to calculate this limiter range. In these equations, x_{RMS} represents the RMS of the input signal (i.e., microphone or accelerometer), and VAR_{factor} is a predefined constant that controls the weight of the RMS variance on the threshold range defined by L_{min} and L_{max}. The variable L_{center} is employed to center the limiter range on an optimal value where all RMS onset and offset events are detectable with the fewest artifacts resulting from the floor of the RMS signal; this mean-to-median ratio minimizes false-positive events that may otherwise occur due to the limiter range being driven by the noise floor. For instance, if instead the median of the RMS signal is used for centering, the limiter range will be lower in individuals who take longer pauses (e.g., during slower vocal rates or when taking a breath). In contrast, centering the limiter range using the L_{center} ratio is a simple way to safeguard against this issue; recordings at slower vocal rates will increase the ratio, thereby raising the limiter range and avoiding faulty onset/offset detections that could appear due to variations in the RMS floor. An example of the event detector is shown in Figure A2, where Figure A2a displays the overlapped microphone and accelerometer signals and Figure A2b identifies the detected events.

VF Estimation Algorithm

After applying the limiting thresholds, a 2-point differentiator is employed to calculate the derivative of the RMS waveform. The standard deviation of the derivative and a predefined tolerance factor are used as parameters to locate strong rises and falls in the derivative waveform, which correspond to voice onsets and offsets in the RMS waveform, respectively. Resulting locations are then evaluated ("sequence validation" in Figure A1) in order to remove nonphysiologically possible rise or fall sequences. This sequence validation process takes the located peaks and a predefined minimal time gap factor between each identified location as inputs to identify potential outliers. Finally, an event resolver distinguishes the remaining time points as either temporal midpoints of a phonatory segment or a segment of vocal fold abduction via a skew factor ratio of the time gap between offset–onset moments. This skew factor is necessary due to the natural delay in oscillatory behavior during offset–onset sequences (Ikuma et al., 2016). Specifically, abduction gestures may be initiated before vocal fold vibration ends, which delays the drop in RMS that occurs during the transition into the voiceless consonant. As a result, the glottic angle waveform rises prior to the end of oscillatory motion. Similarly, vibratory initiation may also be delayed during the transition from the voiceless consonant into the vowel. As such, the skew factor must be chosen as less than 50% of the offset–onset time in order to represent potential time points where maximum abduction may occur. After this process, event tags are assigned to each time point detected.

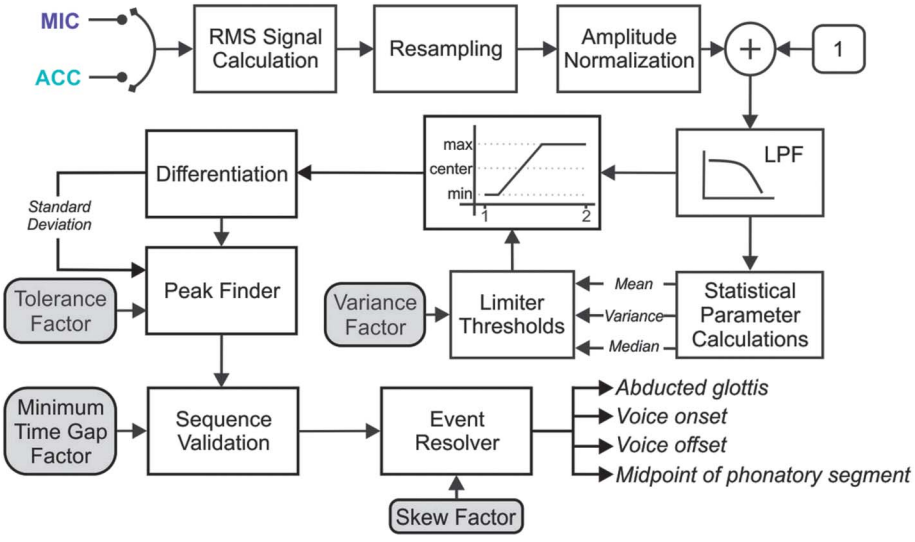
Figure A1 illustrates the event detector procedure that operates on either the microphone or accelerometer signal. The root-mean-square (RMS) of the input signal is first calculated using methodology described by Ikuma et al. (2016), in which the signal bandwidth is restrained to the first three harmonics. Then, the RMS signal is resampled to the frame rate of the HSV recording (i.e., 1,000 fps) and amplitude normalized. Afterward, the signal is low-pass filtered with a first-order Butterworth filter with a cutoff frequency of 25 Hz in order to smooth the RMS responses of short-timed activity on the signal (e.g., glitches, clicks) that are not related to voiced phonation. Mean, median, and variance are computed from this filtered and amplitude-normalized RMS signal to establish limiter thresholds for distinguishing voiced and unvoiced components from silence. Specifically, these thresholds are employed to create a "limiter range" from which events can be distinguished (i.e., abducted vocal folds, voice onset or offset, and vowel midpoint) using the RMS of the auxiliary signal.

$$L_{\text{center}} = \frac{\text{mean}(x_{\text{RMS}})}{\text{median}(x_{\text{RMS}})} \tag{3}$$

$$L_{\text{min}} = L_{\text{center}} - \text{VAR}_{\text{factor}} \cdot \text{var}(x_{\text{RMS}}) \tag{4}$$

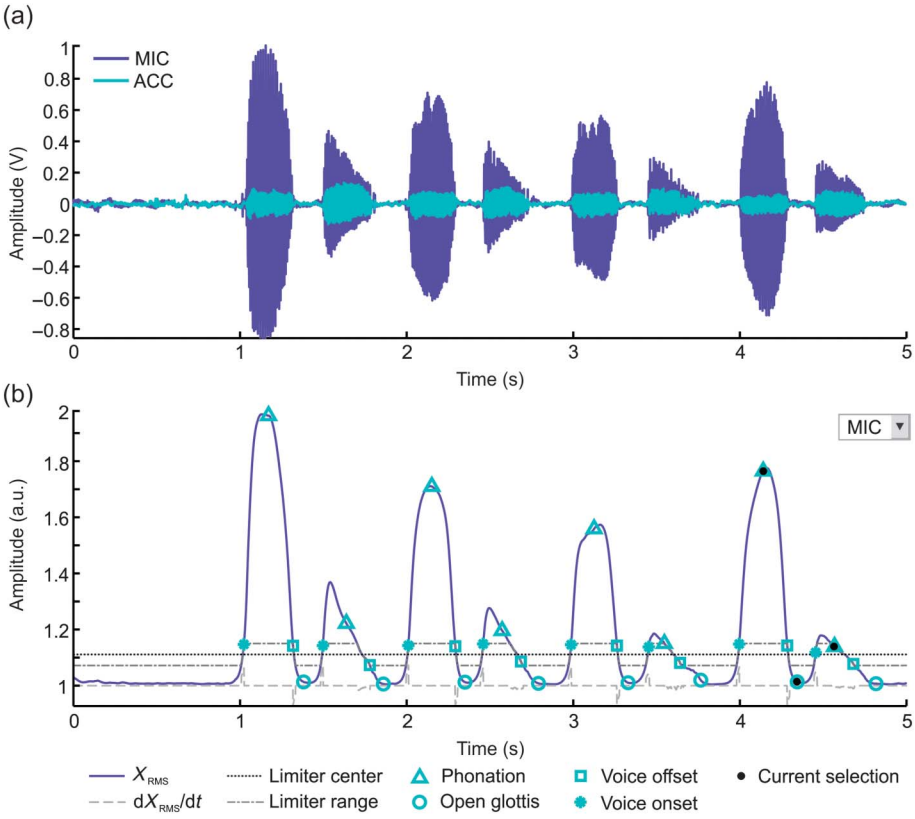
$$L_{\text{max}} = L_{\text{center}} + \text{VAR}_{\text{factor}} \cdot \text{var}(x_{\text{RMS}}) \tag{5}$$

Figure A1. Event detector schematic, with auxiliary (MIC = microphone; ACC = accelerometer) signal as input and event as a function of frame index as output. Additional inputs to the system are highlighted in gray: Tolerance factor, minimum time gap factor, variance factor, and skew factor are each predefined constants. Frames are tagged as a time point corresponding to one of the following events: abducted glottis, voice onset, voice offset, or the temporal midpoint of a phonatory segment.



VF Estimation Algorithm

Figure A2. (a) Example of the microphone (MIC) and accelerometer (ACC) signals obtained from a synchronized high-speed videoendoscopic recording, with three /ifi/ repetitions, and (b) result of the event detection process with the MIC as input. The root-mean-square of the input signal (x_{RMS}; solid purple line) is limited by thresholds L_{min} and L_{max} obtained via signal statistics (shown here as "limiter range" and "limiter center"). The voice onset and offset locations are indicated by the teal asterisk and square markers, respectively. The event resolver indicates instances of phonation (teal triangle markers) at the temporal midpoint of the onset–offset time gaps and unvoiced (i.e., glottis is likely to be open) as teal circle markers at a proportion of the offset–onset time gap. Event selection is controlled by the user and is denoted by the black dots; the location of the three dots defines which part of the video will be selected for further processing.



Preprocessing

Subsequent to event detection, the user is prompted to select a video segment and define initial parameters to begin video analysis. The identified events, in addition to the recorded high-speed images within the chosen segment, serve as input to a user interface (UI) procedure. There are five sequential UI states: (a) video segment selection, (b) glottic axis selection, (c) region of interest (ROI) selection, (d) seed point selection, and (e) segmented region growing (SRG) threshold selection. Figure A3 illustrates glottal axis, ROI, seed point, and SRG threshold selection states. The goal of these preprocessing steps is to construct a set of initial parameters in the automatic glottic angle extraction process that enable the detection and segmentation of the vocal folds.

In order to detect the glottis in each video frame, the user must set a variety of parameters using a reference frame; these parameters enable the initialization of a binary mask that contains all valid positions of the glottis in the remaining video frames. At the start of this binary mask creation process, the UI first displays a time scope window with previously detected events to allow the user to navigate across time events (see Figure A2). The user may move a trio of markers across the window to choose a group of events that correspond to the desired initial, final, and reference processing frames. Initial and final frames are used to extract the video segment, while the reference frame is used to determine an initial ROI, seed points, threshold profile, and glottic axis. With an interest in kinematic offset and onset transition behaviors, the trio of markers should be set to select a single VCV instance; here, the initial and final processing frames correspond to the vowels in the VCV instance, while the reference frame is the voiceless consonant. More specifically, the event selection markers must be placed such that (a) the initial marker is located within the first vowel, (b) the final marker must be located within the latter vowel, and (c) the middle marker ("reference image") must be assigned between the initial and final markers, likely corresponding to a segment where the vocal folds are abducted (i.e., the consonant). An example of a valid three-marker selection is illustrated on Figure A2.

VF Estimation Algorithm

As demonstrated in Figure A3a, the selected reference image must be user-bounded via ROI selection wherein the glottis is centered in the ROI. A glottal axis is drawn to define the extension of vocal folds used during the edge segmentation step and to compensate for the rotated orientation of the scope during the process. After that, seed points corresponding to the glottis may be defined, which then initialize a growing region algorithm over a manipulated grayscale version of the reference image. The growing region is controlled by a vertical threshold profile defined interactively by the user during its selection state. The result of this initialization is a binary mask containing all allowed positions to begin glottis detection in the remainder of the video frames.

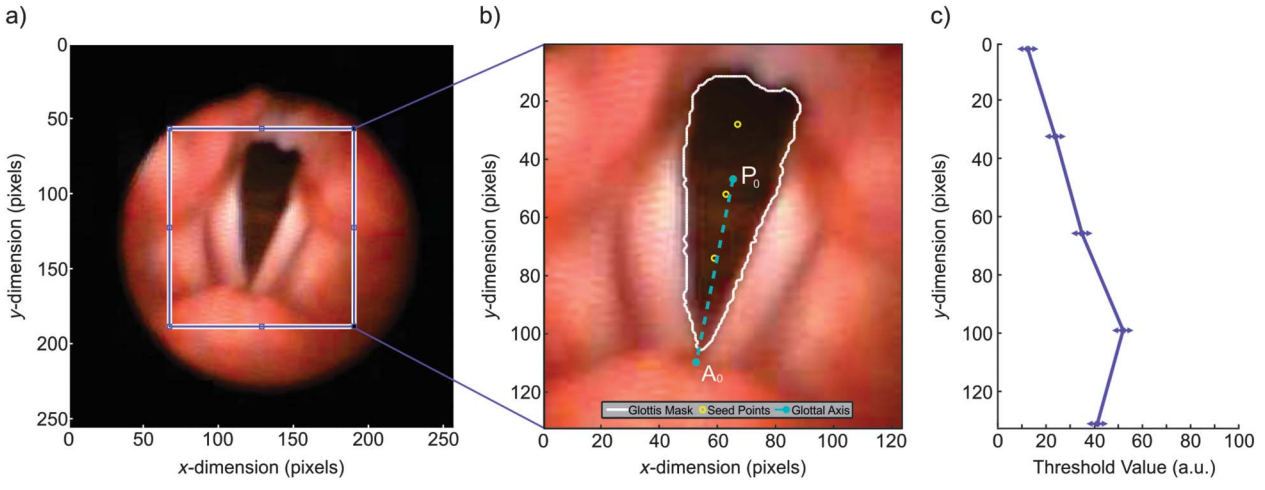
Glottis Detection

Before initializing the glottis detection algorithm (visualized in Figure A3b), the HSV input is first converted to 8-bit grayscale and is manipulated ad hoc in order to reduce potential detection failures under low light or contrast. This process consists of (a) inverted two-dimensional Kaiser window masking, (b) Gaussian edge tapering, and (c) minima imposing. The two-dimensional Kaiser window masking is employed to increase pixel value intensities away from the center of the image, thereby providing a better contrast between the glottal area and dark laryngeal surroundings. Gaussian edge tapering is then applied to smoothly eliminate border information in the ROI. Finally, an operation to impose minima is implemented in order to force pixel intensity values inside the glottal area as minima rather than other local low-intensity regions. The resulting HSV images subsequently undergo glottal segmentation.

The glottal segmentation processing step is crucial for reducing the likelihood of segmentation errors that may occur when pixel intensity values in the glottis are similar to other surrounding dark areas (e.g., due to variation in anatomical structure between individuals or dynamic light orientation of the endoscope throughout the recording session). It is important to note that ROI selection circumvents this issue by ignoring these troublesome regions, but only if the problematic pixels are outside the desired area. Therefore, manipulations are needed when such low-contrast conditions are present between the vocal fold edges.

A modified implementation of the glottal segmentation method proposed by Lohscheller, Toy, Rosanowski, Eysholdt, and Döllinger (2007) is used to identify the glottis from HSV images. This method consists of an SRG algorithm that employs a hysteresis thresholding procedure to segment the glottal area. The result is a binary image containing all pixels that lie within the local limited intensity range as part of the glottis. This range depends on the pixel values of the user-defined glottal mask and the selected threshold corresponding to its vertical pixel location within the user-selected ROI. This method segments the glottis at each border, thereby obtaining two sets of points of raw integer pixel-wise resolution at the boundary between the glottis and each vocal fold ($\overline{C}_{left}, \overline{C}_{right}$). Contrary to the glottal segmentation technique implemented by Lohscheller et al., the current glottis detection algorithm does not update the seed points needed to initialize the detection. Rather than sporadically recalculating the glottal mask during phonation periods, all pixels within the glottal mask area are considered to be seed points during the abduction or adduction gesture.

Figure A3. Preprocessing user interface example, with (a) selected frame as image reference via event selection process shown in Figure A2; (b) result for seed point selection, anterior–posterior glottal axis, and glottal mask; and (c) vertical threshold profile used to compute the glottal mask via segmented region growing algorithms.



One downfall with this glottis detection method, however, is that it does not compensate for barrel distortion or wide-scope positions during adductory movements. It is assumed here that the translational movement of the glottis during adduction is small enough that the glottal mask at least partially overlaps with the glottal area identified inside the adjacent video. In general, this is a fair assumption since the duration of each VCV speech segment is less than 400 ms; however, cases still exist in which the larynx or endoscope point of view (i.e., endoscope depth, distance, and tilt) significantly drifts from the expected location. For those ill-centered angles, errors in glottis detection during glottic angle data extractions are expected and manual glottic angle markings (see Manual Angle Marking) may be useful.

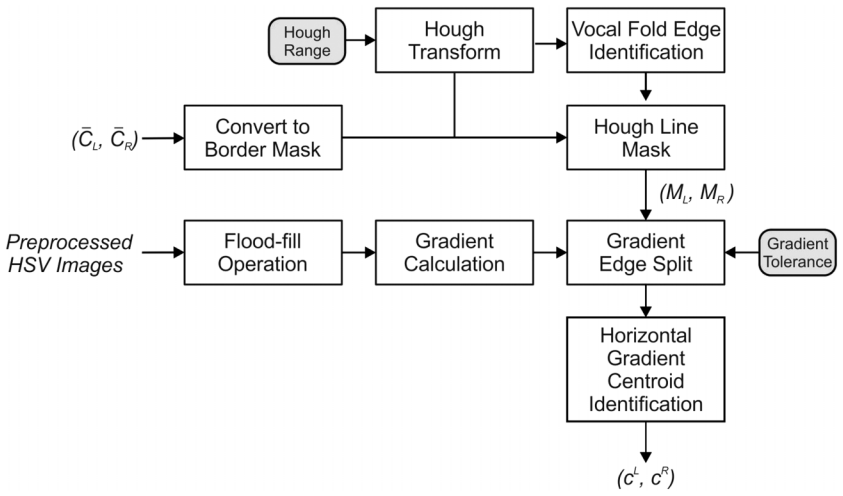
Edge Segmentation

Vocal fold edge segmentation is necessary to (a) remove segments of the point curves produced from the glottal segmentation step that correspond to other laryngeal structures (e.g., arytenoids) and (b) obtain a horizontal subpixel location of each vocal fold border using the resulting gradient information. These borders are the main component of the glottic angle calculation: Subpixel resolution aids in correcting the noisy curves from the glottis detection, instead supplying a smooth curve to analyze angular kinematics from during offset–onset events. Figure A4 shows the edge segmentation step in detail.

First, the two data series corresponding to the glottal borders of the left ($\bar{C}_L$) and right ($\bar{C}_R$) vocal folds are each converted into a binary image. Pixels within each image that correspond to the border between the vocal folds and the glottis are labeled as *true*, and pixels that are attributed to other structures are labeled as *false*. A Hough transform (see Figure A4) is then applied to these images in order to determine the most prominent straight line for the edge data sets. In this step, the Hough transform is restricted to the maximum range of the inclination angle that the Hough representation takes as input for a straight line search; the default magnitude for the Hough range is 90°. The resulting Hough lines are used to prune edge points away from the straightest portion of the vocal fold tissues. The resulting mask images from this process are denoted as M_L and M_R for the left and right vocal fold borders, respectively.

Concurrently, a flood-fill operation based on morphological reconstruction is applied to the processed video frames from the glottis detection step. This operator suppresses highly saturated regions of the image by assigning lower intensities to the local pixel neighborhood. The spatial gradient (magnitude and phase) is computed by convolving the resulting video frames with a Prewitt operator. As a result, gradient values of saturated regions are eliminated to prevent potential issues that could affect vocal fold edge calculation. The gradient magnitude is then split according to left and right vocal fold edges via the M_L and M_R masks and the gradient phase information. Within this step, a gradient tolerance (see "Gradient Tolerance" in Figure A4) is applied to establish the minimum gradient magnitude accepted as gradient information in the image; the default magnitude of this parameter is 0.2. Equations 6 and 7 show the calculation for the splitting masks of the right vocal fold (B_R) and left vocal fold (B_L) that arise from the gradient masks. If the gradient magnitude is G and its phase is G_Φ, then the splitting process follows Equations 8 and 9 to determine the split gradients for the right vocal fold (G_R) and left vocal fold (G_L).

Figure A4. Edge segmentation overview, with left and right vocal fold boundaries ($\bar{C}_L$, $\bar{C}_R$, respectively) taken as inputs to the system. Left and right vocal fold boundaries are each transformed into a gradient mask (M_L and M_R, respectively) using the Hough transform. These masks are then applied to the preprocessed HSV gradient to split the gradient magnitude into left and right vocal fold edges. These edges are then used to calculate horizontal subpixel values of each vocal fold edge (c^L, c^R).



$$B_{\mathrm{R}} = \begin{cases} M_{\mathrm{R}}, & G_{\Phi} > 90 \vee G_{\Phi} < -90 \\ 0, \text{otherwise} \end{cases} \tag{6}$$

$$B_{\mathrm{L}} = \begin{cases} M_{\mathrm{L}}, & G_{\Phi} < 90 \wedge G_{\Phi} > -90 \\ 0, \text{otherwise} \end{cases} \tag{7}$$

$$G_{\mathrm{R}} = \begin{cases} G \cdot B_{\mathrm{R}}, G > G_{\mathrm{TOL}} \\ 0, \text{otherwise} \end{cases} \tag{8}$$

$$G_{\mathrm{L}} = \begin{cases} G \cdot B_{\mathrm{L}}, G > G_{\mathrm{TOL}} \\ 0, \text{otherwise} \end{cases} \tag{9}$$

Once the split gradients are calculated, centroid values are computed, as shown in Equation 10, in order to localize the horizontal subpixel values of each vocal fold edge (left, L; right, R) along the vertical axis.

$$c^s : (x^s, y^s) = \left(\frac{\sum_{i=1}^{w} i \cdot G_s(i,j)}{\sum_{i=1}^{w} G_s(i,j)}, j \right) \quad \text{where } \forall j \in [1,h], s \in [\mathrm{L}, \mathrm{R}] \tag{10}$$

Variables w and h correspond to the width and height of the ROI image, respectively. It is important to note that vertical subpixel values for both vocal fold edges preserve the y-axis index (j). Finally, the sets c^s are extended by two additional points: the anterior (A_0) and posterior (P_0) points from the defined glottal axis in the preprocessing module.

Glottic Angle Estimation

A line-fitting process to estimate the glottic angle waveform is performed once the edge curves c^{L} and c^{R} are calculated. Using slope and anterior commissure location restrictions, a pair of vocal fold lines are computed using the edge curves via Equation 11, where Φ^{L} and Φ^{R} intersect at a single point that corresponds to the anterior commissure. This point is denoted as $\bar{v} = (\bar{v}_x, \bar{v}_y)$ and is restricted to the range described in Equation 12.

$$\Phi^s = (m^s, b^s) \tag{11}$$

$$\bar{X}_1 \le \bar{v}_x \le \bar{X}_2, \bar{Y}_1 \le \bar{v}_y \le \bar{Y}_2 \tag{12}$$

We denote these ranges as $\bar{X}$ and $\bar{Y}$, where $\bar{v}$ has a bounding line solution property restricting solutions only where their intercept lays around a physiologically plausible location in the image in terms of vertical subpixels $\bar{X}$ and horizontal subpixels $\bar{Y}$. However, these boundaries are calculated slightly differently depending on whether manual intervention was used. When the algorithm runs without any manual aid, the range of $\bar{X}$ is fixed and centered onto the horizontal subpixel location corresponding to wherever the anterior point was selected during the preprocessing step; however, the range of $\bar{Y}$ is not centered and depends on the median position of both edge curves (c^s), as shown in Equations 13 and 14.

$$\bar{X} = [A_{0x} - 0.025L_0, \; A_{0x} + 0.025L_0] \tag{13}$$

$$\bar{Y} = \left[\mathrm{median}\left(c_y^{\mathrm{L}}\right), \mathrm{median}\left(c_y^{\mathrm{R}}\right) \right] \tag{14}$$

Here, $A_0 = (A_{0x}, A_{0y})$ is defined as the anterior point that was fixed during preprocessing steps (see Figure A3b, only the x-axis component is considered) and L_0 is the length of the selected glottal axis. With this setup, lateral corrections for the anterior point are possible when there are left–right shifts in c^s with respect to the reference image, yet vertical estimates of anterior commissure $\bar{v}_x$ are confined to the fixed range defined in Equation 13.

Once the restrictions for $\bar{v}$ are established, Φ^L and Φ^R can be optimized according to Equations 15–19. Here, the objective function (see Equation 15) is minimized by reducing perpendicular errors between the edge curves c^s and the line Φ^s in a Geman–McClure sense with the robust function, $\rho(e)$.

$$\left(\tilde{\Phi}^L, \tilde{\Phi}^R\right) = \arg \min_{\tilde{\Phi}^L, \tilde{\Phi}^R} \frac{1}{N_R + 2} \sum_{i=1}^{N_R} \rho\left(E\left(c^R(i), \Phi^R\right)\right) + \frac{1}{N_L + 2} \sum_{i=1}^{N_L} \rho\left(E\left(c^L(i), \Phi^L\right)\right) + \lambda\, \rho\left(0.5\bar{Y}_1 + 0.5\bar{Y}_2 - \bar{v}_y\right) \quad (15)$$

where $\bar{X}_1 \le \bar{v}_x \le \bar{X}_2,\ \ \bar{Y}_1 \le \bar{v}_y \le \bar{Y}_2, -1 \le m^s \le 1$

$$\Phi^s = (m^s, b^s), s \in [L, R] \quad (16)$$

$$E(\varsigma, \Phi^s) = \frac{|m^s \varsigma_x + b^s - \varsigma_y|}{\sqrt{(m^s)^2 + 1}} \quad (17)$$

$$\rho(e) = \frac{e^2}{(\sigma^2 + e^2)}, \sigma = 5 \quad (18)$$

$$\bar{v} = \left(\frac{m^R b^L - m^L b^R}{m^R - m^L}, \frac{b^L - b^R}{m^R - m^L}\right) \quad (19)$$

Since the edge curves include anterior A_0 and posterior P_0 points, the number of detected points on each edge is increased by 2. Thus, the terms will never be undetermined. A third term is included to drive valid $\bar{v}$ solutions within lateral range. As well, two additional restrictions are included to limit the possible maximum angle of the intersecting slopes of $\bar{v}_x$ and $\bar{v}_y$ to below 90°. This optimization is performed via sequential quadratic programming for each HSV frame, in which the initial condition is calculated using common linear least-squares solutions over c^s. In the event of very narrow or strongly adducted vocal fold edges, two vertical lines are instead assigned as the solution, with each line located over the previously calculated anterior commissure $\bar{v}$. Narrow edges are empirically defined as if the $\bar{Y}$ range is less than a quarter of a pixel. Therefore, the final vocal fold edge line estimates, $\hat{\Phi}^s$, are computed as in Equation 20.

$$\hat{\Phi}^s = \begin{cases} \tilde{\Phi}^s, & \hat{Y}_2 - \hat{Y}_1 > 0.25 \\ (0, \bar{v}_y), & \text{otherwise} \end{cases} \quad (20)$$

Finally, the glottic angle, θ, between vocal fold lines is calculated using the slopes of the vocal fold edges, $\hat{m}^s$ using the trigonometric property defined in Equation 21.

$$\theta = \arctan\left(\frac{\hat{m}^L - \hat{m}^R}{1 + \hat{m}^L \hat{m}^R}\right) \quad (21)$$

Manual Angle Marking

Trained users may opt for manual intervention during the glottic angle extraction procedure, specifically in instances where the user does not agree with the algorithmic results. In situations where the algorithmic estimations (see Figure 2 for an example of the algorithm output) fail to appropriately track the glottic angle over time, the user can manually mark glottic angles to use as input to the algorithms.

The user-defined parameters set during the threshold profile and glottis mask selection states are considered as constant parameters in the remaining processing stages under the assumption that large variations in glottal position during adduction will not occur. However, the glottal axis location can be manipulated by a manual angle marking procedure if the user finds that the glottis is considerably displaced with respect to the reference frame. Specifically, if the previously defined anterior/posterior location is not satisfactory for all abduction and adduction sequences, the user may incorporate manual angle markings following methodology by Dailey et al. (2005) on video sequences down-sampled to 50 fps. Following manual marking, the remaining angles not estimated in the down-sampled signal are linearly interpolated to the video sequence at 1,000 fps. Here, the algorithm uses these manual markings as a guide to recalculate and place new anterior and posterior points; it is recommended that this manual analysis scheme is used when noticeable glottal translations occur. When manual glottic angle data are used, the intercept between their lines establish a manually defined anterior point $A_m = (A_{mx}, A_{my})$ that now varies across time. These new coordinates replace the anterior point $A_0 = (A_{0x}, A_{0y})$ that was defined during preprocessing steps. Both $\bar{X}$ and $\bar{Y}$ regions are designed as a function of A_m, as shown in Equations 22 and 23.

$$\bar{X} = [A_{mx} - 0.025L_0,\ \ A_{mx} + 0.025L_0] \tag{22}$$

$$\begin{aligned} &\bar{Y} = \left[A_{my} - 0.5d_{\text{GAP}},\ \ A_{my} + 0.5d_{\text{GAP}}\right] \\ &\text{where } d_{\text{GAP}} = \left|\,\text{median}\left(c_y^{\text{L}}\right) - \text{median}\left(c_y^{\text{R}}\right)\,\right| \end{aligned} \tag{23}$$

Algorithm Validity

In order to determine the validity of the glottic angle extraction algorithm, the smooth glottic angle data resulting from the algorithm were directly compared to the manual glottic angle data. For this comparison, two individual /ifi/ productions were randomly selected from each participant, with one production selected from a speed condition (i.e., SR, RR, FR) and one production selected from an effort condition (i.e., MIL, MOD, MAX). These productions were extracted from 69.8% of cases wherein technicians accepted the initial automated angle results. Cases that were accepted following manual aid or were considered unusable were not considered for assessing algorithm validity in order to directly compare the initial automated results of the algorithm against corresponding manual angle estimates. Manual angle markings were performed on each selected production at a down-sampled rate of 100 fps by two trained technicians who were blind to the data set. Table A1 shows the two-way intraclass correlation coefficients (ICCs) for consistency computed between each trained technician and the automated algorithm, as well as between technicians. The technicians performed with good-to-excellent reliability (Koo & Li, 2016) when compared to each other (ICC = .89 with 95% CI [.86, .91]). Moreover, the algorithm performed with good reliability when compared with the manual markings of each technician, yielding ICC = .82 (95% CI [.77, .86]) with the first technician and ICC = .84 (95% CI [.80, .88]) with the second technician. A final two-way ICC analysis for consistency was then computed to compare the results of the algorithm with the average of the technicians' manual angle markings. This analysis yielded good reliability between the glottic angle estimates resulting from the automated algorithm and those from manual angle markings, with ICC = .85 (95% CI [.81, .89]).

Table A1. Intraclass correlation coefficients (ICCs) and 95% confidence intervals (CIs) between each trained technician and automated algorithm as well as between the trained technicians.

Comparison	ICCs (95% CI)
Technician 1 × Automated Algorithm	.82 [.77, .86]
Technician 2 × Automated Algorithm	.84 [.80, .88]
Technician 1 × Technician 2	.89 [.86, .91]
Averaged Technicians × Automated Algorithm	.85 [.81, .89]