Published in final edited form as:

J Voice. 2017 May ; 31(3): 389.e9–389.e17. doi:10.1016/j.jvoice.2016.09.021.

Relationship between acoustic voice onset and offset and selected instances of oscillatory onset and offset in young healthy males and females

Rita Patel, Ph.D., CCC-SLP, Karen Forrest, Ph.D., and Drew Hedges, B.S.

Department of Speech and Hearing Sciences, Indiana University, Bloomington, Indiana

Abstract

Objective—To investigate the relationship between (1) onset of the acoustic signal and pre-phonatory phases associated with oscillatory onset and (2) offset of the acoustic signal with the post-phonatory events associated with oscillatory offset across vocally healthy adults.

Subjects and Methods—High-speed videoendoscopy was captured simultaneously with the acoustic signal during repeated production of /hi.hi.hi/ at typical pitch and loudness from 56 vocally healthy adults (age 20–42 years; 21 male, 35 female). The relationship between the acoustic sound pressure signal and oscillatory onset/offset events from the glottal area waveforms (GAW), were statistically investigated using a multivariate linear regression analysis.

Results—The onset of the acoustic signal (X_{1a}) is a significant predictor of the onset of first oscillations (X_{1g}) and onset of sustained oscillations (X_{2g}). X_{1a} as well as gender are significant predictors of the first instance of medial contact ($X_{1.5g}$). The offset of the acoustic signal (X_{2a}) is a significant predictor of the first instance of oscillatory offset (X_{3g}), first instance of incomplete glottal closure ($X_{3.5g}$), and cessation of vocal fold motion (X_{4g}).

Conclusions—The acoustic signal onset is closely related to the first medial contact of the vocal folds but the latency between these events is longer for females compared to males. The offset of the acoustic signal occurs immediately after incomplete glottal adduction. The emerging normative group latencies between the onset/offset of the acoustic and the GAW from this study appear promising for future investigations.

Keywords

Oscillatory onset time; voice onset time; oscillatory offset time; voice offset time; high-speed videoendoscopy; sound pressure; acoustic

Author to whom correspondence, reprint requests, and proofs should be sent: **Rita R. Patel, Ph.D., CCC-SLP**, Indiana University, Dept. of Speech & Hearing Sciences, 200 S. Jordan Avenue, C145, Bloomington, IN 47405-7002, patelrir@indiana.edu, Phone: 812-855-3886, Fax: 812-855-5561.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Conflict of Interest: None

Financial Disclosure: None

INTRODUCTION

The relationship between onset and offset of vocal fold oscillation and the start and end of the acoustic signal has been a longstanding topic of interest in the field of speech science. Numerous studies have investigated the factors that contribute to the estimation of timing and coordination between the laryngeal and supralaryngeal systems by investigating the voice onset time¹ from the acoustic signal during the production of voiceless consonants.^{1–4} Voice onset time is defined as the interval between the release of an oral constriction of a stop consonant and the start of the vocal fold vibrations for the following vowel.¹ Studies evaluating the voice onset time predominantly determine the onset of vocal fold vibrations from spectrographic or waveform analysis of the acoustic signal. Very few studies have focused on examining the true oscillatory onset and offset behavior at the level of the vocal folds with techniques of laryngeal imaging and its relationship with the onset and offset of the periodic acoustic signal. The relationship between the source and the filter coupling at onset and offset is not only important for classifying different types of onsets (e.g. soft, breathy, hard), but also for investigating the effects of non-stationary vibratory motion on the acoustic output. Empirical investigations into the precise timing between the source and the filter characteristics can provide insights into factors like subglottal pressure, glottal size, vocal fold length, stiffness, phonation threshold pressure, glottal adduction that control onset and offset of voicing.^{5,6}

With the advancement of digital technology and the clinical availability of high-speed videoendoscopy^{7–9} and videokymography¹⁰ for directly imaging the vocal fold oscillations, it is possible to capture individual vocal fold cycles and transient behaviors associated with voice onset and offset. The literature on direct studies of vocal fold oscillations with *simultaneously* recorded acoustic signal is exploratory at best and limited to analysis of one to two subjects. With the development of a multidimensional voice analysis system Köster et al (1999)⁷ demonstrated large intrasubject variations in the pre-phonatory phase in one female subject in two repeated productions of /ihi/. Using simultaneous high-speed videoendoscopy, electroglottography, and acoustic recordings in one Korean speaker, Hong et al (2002)¹¹ reported the presence of the largest glottal gap for the heavily aspirated bilabial consonant, a somewhat smaller gap for aspirated, and smallest gap for unaspirated tokens, resulting in the highest intensity of aspiration noise and longer voice onset time for the heavily aspirated type. McDonnell et al (2011)¹² evaluated the onsets of the glottal adduction during staccato, aspirated, and unaspirated production of /pi/ using simultaneous high-speed video and spectrographic analysis of the acoustic signal in 6 trained singers and found that the initiation of vocal fold contact and appearance of high frequency ripple in the audio signal occur at a later time in aspirated compared to the unaspirated and the staccato productions. We were unable to locate sources in the literature that examined direct vocal fold oscillatory motion with simultaneous acoustic recording of voice offset or any large-scale studies relating pre-phonatory vocal fold activity to the onset of the acoustic signal.

Limited studies investigating oscillatory onset using high-speed videoendoscopy and videokymography have demonstrated the presence of the following pre-phonatory phases: (1) vocal fold adductory phase, (2) onset of small irregular oscillatory motion of the vocal fold margins, (3) first instance of medial contact, and (4) onset of steady-state cyclic

oscillations.^{8,12–18} Similarly, oscillatory offset has been demonstrated to broadly consist of the following post-phonatory phases: (1) brief duration of non-steady state oscillatory motion typically involving an increased open phase and reduction in oscillatory amplitude, followed by (2) complete cessation of oscillatory motion and vocal fold abduction to rest breathing.⁸ In the current study particular interest is given to examining the relationship between the various pre-phonatory and post-phonatory events associated with oscillatory onset and offset to the start and termination of the acoustic signal.

The relationship between vocal fold contact and the acoustic pressure signal has been investigated by comparing the acoustic and the electroglottographic (EGG) signals to derive the Vocal Attack Time (VAT).^{19,20} The VAT was defined by Orlikoff et al (2009)²⁰ as the time lag between the rise of the sound pressure in the acoustic signal and EGG signal at the onset of phonation. The measure of VAT was validated in 5 vocally normal subjects during breathy, normal, and hard onset by creating a digital kymogram from high-speed video endoscopy and comparing the VAT during simultaneous EGG and sound pressure signal measurements.²⁰ Subsequently Roark et al (2012)¹⁹ conducted a follow up large scale EGG study (females = 57, males = 55) examining the age (20–50 years) and gender differences in the VAT for three tokens of the vowel /a/ at comfortable pitch and loudness. The mean VAT values were significantly shorter for females compared to males; however, there was no systematic effect of age. Roark et al (2012)¹⁹ speculated that the shorter VAT values in females may be related to smaller laryngeal dimensions, shorter vocal folds, and/or increased stiffness of the vocal folds in females.¹⁹ Similar to the VAT, Watson et al (2012)²¹ explored and defined voice offset in adults using the acoustic and EGG recordings by establishing the measure of Vocal Release Time (VRT). The vocal release time was defined as the lag time between the termination of the sound pressure and EGG signals. Overall, the mean lag times of VRT²¹ were comparatively longer than VAT measures.¹⁹ Contrary to the VAT, the mean VRT values were significantly longer for females compared to males. The authors offered no explanation for the gender differences between the VRT other than the lack of linguistic constraints for offset.²¹ Though useful, the EGG offers only indirect measures of vocal fold contact. The limitations of using EGG signal to obtain vocal fold contact measures include the validity of such vibrations due to potential mucous obstruction, which may hinder the estimations. Further, pre-phonatory oscillations that occur before vocal fold contact could not be captured well by EGG.

Because acoustic analysis is a frequently used, noninvasive method to derive inferences regarding the oscillatory onset/offset, our goal here is to evaluate the relationship between the acoustic onset/offset with the oscillatory onset/offset. The purpose of this study was to evaluate the timing relationship between the onset and offset of the acoustic waveform with the various oscillatory onset and offset events in the glottal area waveform (GAW) obtained from high-speed videoendoscopic recordings of males and females with typical voice. More specifically the relationship between the onset of the acoustic waveform was compared with the following three oscillatory onset related time events in the glottal area waveform: (1) onset of first oscillatory motion, (2) first medial contact of the vocal fold, and (3) start of sustained oscillations in the context of /hi/ productions across males and females. Similar to onset, the relationship between the offset of the acoustic waveform was compared with the following three oscillatory offset related time events in the glottal area waveform: (1) first

instance of sustained oscillatory offset (departure from sustained phonation), (2) first instance of incomplete medial vocal fold contact, and (3) complete cessation of oscillatory motion. For onset measures, we predict that the vocal folds will begin oscillations before the onset of the acoustic signal and will reach a steady state of phonation well after the onset of acoustic energy. Regarding offset measures, we predict that the acoustic signal will end before vocal fold oscillation cessation, but after the end of steady-state phonation.

METHODS

Participants

A total of 92 adults (ages 20–45 years) with no previous vocal fold pathologies were recruited at the University of Kentucky's Vocal Physiology and Imaging Laboratory and signed of an Institutional Review Board approved informed consent form. All recruited participants were rated perceptually to have normal voice by a certified speech language pathologist on the GRBAS scale,²² were not professional voice users, and had negative history of smoking. Due to heightened gag reflex data from 21 subjects could not be obtained. The current report includes data from 56 (age 20–42 years; 21 male, 35 female) subjects whereas other data were excluded due to the following limitations: failure to extract the audio signal ($n = 10$) and failure to synchronize the video and audio signals ($n = 5$).

A minimum of three consecutive repetitions of /hi/ syllable productions at self-selected typical conversational pitch and loudness levels were recorded. The average fundamental frequency in females was 242.16 ± 27.66 Hz and for males was 153.09 ± 24.68 Hz. The average loudness in females was 64.71 ± 5.56 dB SPL and for males was 61.94 ± 5.72 dB SPL. The combination of a voiceless fricative /h/ and a vowel /i/ offers an ideal method of assessing oscillatory onset and acoustic voice onset simultaneously because /h/ is produced with a fairly open tract without supraglottal constrictions. The voiceless fricative /h/ has also been demonstrated to play a neutral role on the subsequent vowel production,^{2,3} thereby obtaining the true physiological onset at the level of the vocal folds without complicating effects of the supraglottal constrictions and potentially preventing a hard glottal attack.⁷ A 70-degree rigid endoscope was used to capture the high-speed videoendoscopic recording of the vocal fold oscillations at 4000 frames per second. High-speed videoendoscopic recordings were obtained using PENTAX Digital Model 9710 (Montvale, New Jersey) with a spatial resolution of 512×256 pixels. A microphone was placed at a mouth-to-microphone distance of 5 inches to capture the acoustic signal at a sampling rate of 50kHz simultaneously with the high-speed videoendoscopic recordings.

Data Segmentation

The middle syllable production of /hi/ was segmented from the high-speed videos and the co-occurring acoustic record for all subjects to prevent any unusual abrupt occurrences associated with the beginning and termination of phonation. A custom developed image segmentation tool called the Vocal Cord Analyzer (VCA) beta software²³ was used to segment the corresponding video and audio file for the middle /hi/ production. To assure the onset and offset behavior during /hi/ were captured within the segmented window, the file start frame was selected while the arytenoids were in a fully abducted position, immediately

before the first instance of adductory motion, and the end frame of the video was selected when the arytenoids were in the fully abducted position immediately after oscillatory motion of the vocal folds ceased. The synced acoustic signal also was segmented using these landmark behaviors of the glottal wave.

Preprocessing of the video and the audio signal

GAW representing the oscillatory onset and offset behavior of the vocal fold was extracted from the high-speed video using the, Glottal Analysis Tools, (GAT)²⁴ software application, developed at the Department for Phoniatrics and Pediatric Audiology, Erlangen, Germany. The PRAAT²⁵ software was used for display of the acoustic signal. Excess noise in the range of 80–10,000 Hz was removed from the acoustic signal with the Boll's (1979)²⁶ spectral subtraction noise reduction method as implemented in PRAAT. The glottal area waveform and the filtered acoustic waveform for each subject were subsequently time aligned using MATLAB software (Mathworks Inc.). A synchronization pulse from the PENTAX Digital system Model 9710 first registered the acoustic and video recordings (Montvale, New Jersey), where this pulse occurred exactly on a video frame recording. When the recordings were created with the PENTAX Digital system (Montvale, New Jersey), a Camera Information Header data file was automatically generated that included the acoustic synchronization pulse delay, the video frame rate, and the acoustic sampling rate. The MATLAB algorithm rescaled the acoustic signal, recorded at a higher sampling rate than the glottal wave and therefore had more samples, to fit into the same temporal window as the glottal waveform. These time-aligned signals were displayed on the computer screen and the user selected the onset and offset points of each signal. As described below, the experimenter made additional temporal markings to denote the onset and offset events of interest. Repeated measures of each interval were strongly related with Pearson correlations above 0.9 for all measures.

To evaluate the timing relationship between the onset and offset events of the vocal fold oscillations and the acoustic signal, eight distinct measures (3 for oscillatory onset, 3 for oscillatory offset, and 2 for acoustic onset and offset) were manually selected. The following measures were made for each /hi/ production (Figure 1) and were measured in milliseconds relative to the time of onset of the file:

1. **Onset of first oscillatory motion (X_{1g}):** was defined as the duration between the onset of the file and the first detectable small amplitude motion in the GAW during the pre-phonatory phase.¹⁵
2. **First medial contact of the vocal folds ($X_{1.5g}$):** was defined as the start of mid-membranous contact between the left and the right vocal folds as observed on the high-speed video.²⁷ Previous studies have used the EGG signal to estimate vocal fold contact.^{19,20,28} Using high-speed videoendoscopy, we can visually detect the exact time reference of the first medial contact, independent of the amount of contact, in relation to the acoustics more directly.
3. **Onset of sustained oscillations (X_{2g}):** was defined as the first instance in which the GAW reached a steady state oscillation in terms of amplitude periodicity, time periodicity, and glottal closure.^{15,29,30} The closed phase prior to the instance

in which the GAW reached steady state in terms of amplitude periodicity, time periodicity, and glottal closure was selected.

4. **First instance of oscillatory offset (X_{3g}):** was defined as the first departure from the steady state oscillation in terms of amplitude periodicity, time periodicity, and glottal closure in the GAW. The last point at which the vocal folds are closed before the cessation of steady state oscillation was selected.^{30,31}
5. **First incomplete vocal fold closure ($X_{3.5g}$):** was defined as the initial time of incomplete closure between the vocal folds as observed on the high-speed video. An incomplete closure was defined as a glottal gap along the entire length of the vocal folds.^{32,33}
6. **Complete cessation of oscillatory motion (X_{4g}):** was defined as the time following the end of the last detectable non-zero amplitude of the GAW.
7. **Onset of the acoustic signal (X_{1a}):** The onset of the acoustic voicing signal was defined by the point in the waveform associated with the first periodic deviation from the zero baseline of the filtered acoustic signal relative to the onset of the analysis file.
8. **Offset of the acoustic signal (X_{2a}):** The time point after the last detectable non-zero amplitude of the acoustic waveform.

Reliability of Measurements

A total of 15 participants (10 females and 5 males) were selected randomly for analysis of inter-rater and intra-rater reliability. The Pearson product moment correlations and absolute difference for each measure were calculated between the initial and subsequent measurement for each of the variables. The Pearson product moment correlations for intra-rater reliability for X_{1a} , X_{2a} , X_{1g} , $X_{1.5g}$, X_{2g} , X_{3g} , $X_{3.5g}$, and X_{4g} exceeded 0.97 for all variables with an absolute difference for intra-rater comparisons of X_{1a} , X_{2a} , X_{1g} , $X_{1.5g}$, X_{2g} , X_{3g} , $X_{3.5g}$, and X_{4g} of 3, 3, 4, 1, 4, 7, 2, 2 milliseconds. Similarly, the Pearson product moment correlations for inter-rater reliability X_{1a} , X_{2a} , X_{1g} , $X_{1.5g}$, X_{2g} , X_{3g} , $X_{3.5g}$, and X_{4g} ranged from 0.93 to 0.99 for all measures, and the absolute differences for inter-rater comparisons of X_{1a} , X_{2a} , X_{1g} , $X_{1.5g}$, X_{2g} , X_{3g} , $X_{3.5g}$, and X_{4g} were 3, 5, 6, 2, 7, 12, 5, 4 milliseconds.

Statistical Analysis

The emphasis of the current study is on the association between the acoustic and the oscillatory vocal fold signals. Specifically we sought to answer the question of whether or not the timing from the GAW could be reliably determined from knowing only the onset and offset of acoustic signals. Given the presence of multiple variables associated with oscillatory events and the high degree of correlation between them (Table 1), a multivariate linear regression analysis was conducted. Multivariate linear regression is an extension of ordinary multiple linear regression and posits the following model:

$$Y_{(n \times d)} = X_{n \times (p+1)} \beta_{(p+1) \times d} + \epsilon_{n \times d}$$

where d is the number of outcome variables (6), n is the sample size (56), and p is the number of predictors ($3 - X_{1a}, X_{2a}, \text{Gender}$). The use of this technique allows for correlations to exist between the dependent variables represented in the $\mathbf{Y}$ matrix. Our model was constructed to simultaneously regress the three oscillatory onset instances ($X_{1g}, X_{1.5g}, X_{2g}$) onto the acoustic onset of voicing (X_{1a}), the three oscillatory offset instances ($X_{3g}, X_{3.5g}, \text{and } X_{4g}$) onto the acoustic offset of voicing (X_{2a}), and all oscillatory instances ($X_{1g}, X_{1.5g}, X_{2g}, X_{3g}, X_{3.5g}, \text{and } X_{4g}$) onto gender. This model was fit using Mplus version 7.3 (Los Angeles, California). The p values were adjusted for multiple comparisons using the Bonferroni procedure.

Results

All model estimations were carried out using maximum likelihood and using 1000 bootstrapped samples. Our model showed excellent fit statistics, suggesting that the specified model fits the data well. All data fit within this model and, despite some variability across subjects, no individual was a true outlier as evidenced by the fit of all subjects along the linear regression line. The Chi-Square (χ^2) test against a saturated model was statistically non-significant, indicating that our model predicted as well as a saturated model ($\chi^2(6) = 8.31, p = 0.22$). The Root Mean Square Error of Approximation (RMSEA) of our model was estimated at 0.08 with a confidence interval of (0.00, 0.21), which includes the 0.05 threshold determining a quality fit. Other fit statistics like the Confirmatory Fit Index (CFI), the Tucker Lewis Index (TLI), and the Standardized Root Mean Square Residual (SRMR) were well above and below respective conventional standards of quality (CFI = 0.99, TLI = 0.99, SRMR = 0.01). In addition to a good fitting model, we found that the onset of the acoustic signal (X_{1a}) was a significant predictor for all instances of oscillatory onset ($X_{1g}, X_{1.5g}, X_{2g}$) (Figure 2). The following three equations were found to best estimate each instances ($X_{1g}, X_{1.5g}, X_{2g}$) of oscillatory onset:

$$X_{1g} = -20.49 + 1.013 * x_{1a} + 3.67 * \text{Gender}$$

$$X_{1.5g} = 3.26 + 0.991 * x_{1a} + 5.22 * \text{Gender}$$

$$X_{2g} = 18.32 + 1.072 * x_{1a} + 10.26 * \text{Gender}$$

The mean onset of the acoustic signal (X_{1a}) was found to be 79.16 ms in males and 89.96 ms in females relative to the onset of the analysis file, and the median was found to be 67.83 ms in males and 69.84 ms in females. Based on the respective gender mean and median values of the onset of the acoustic signal, our model predicts the mean and median of the timing for onset of first oscillatory motion (X_{1g}) to be 59.66 ms and 48.22 ms, respectively for males. For females the mean and median times for (X_{1g}) are 74.31 ms and 53.93 ms respectively. Similarly, timing of the first medial contact of the vocal folds ($X_{1.5g}$) for males at their mean and median times are estimated at 81.67ms and 70.48ms, respectively. For females the mean and median times for ($X_{1.5g}$) were estimated as 97.63ms and 77.69ms respectively. Lastly,

the estimated mean and median timings of the beginning of sustained oscillations (X_{2g}) for males were found to be 103.15ms and 91.03ms, respectively. For females, the mean and median times for (X_{2g}) are estimated to be 125.02ms and 103.45ms. Thus we find that on average the onset of first oscillatory motion (X_{1g}) occurs before the onset of the acoustic signal (X_{1a}) ($p = 0.001$); onset of the acoustic signal occurs very near, but on average earlier than the timing of the first contact of the vocal folds ($X_{1.5g}$) ($p = 0.012$); and timing of sustained oscillation (X_{2g}) occurs on average after the onset of the acoustic voicing signal (X_{1a}) ($p = 0.012$) (Figure 3). The difference between genders is statistically significant only for the timing of first contact ($X_{1.5g}$). Our model predicts that on average the time between the acoustic onset and the first time of vocal fold contact ($X_{1.5g}$) is 5.22ms ($p = 0.012$) longer for women than for men. Gender differences were not observed for the onset of the first instance of oscillatory motion (X_{1g}) ($p = 1.000$) or the start time of sustained oscillations (X_{2g}) ($p = 0.324$) (Table 2). As such, the gender factor in the model assumes a value of zero for non-significant observations.

Our model also found a significant relationship between the timing of the offset of the acoustic signal (X_{2a}) and all three measures of offset of oscillatory motion (X_{3g} , $X_{3.5g}$, and X_{4g}) (Figure 2). The following three equations were found to best estimate our three time points of oscillatory offset:

$$X_{3g} = -36.67 + 0.94 * x_{2a} - 12.63 * \text{Gender}$$

$$X_{3.5g} = -0.657 + 1.005 * x_{2a} - 7.21 * \text{Gender}$$

$$X_{4g} = 46.26 + 0.992 * x_{2a} - 11.11 * \text{Gender}$$

The mean of the offset of the acoustic signal (X_{2a}) was found to be 305.46ms in males and 318.67ms in females, and the median was found to be 311.6ms in males and 309.7ms in females. Based on the gender respective mean and median values of the offset of the acoustic signal our model estimates the timing of the first instant of departure from sustained oscillatory motion (X_{3g}) for male speakers as 250.77ms and 256.54ms. For females we estimate the mean and median of (X_{3g}) to be 250.55ms and 242.12ms respectively. Similarly, the estimated timing of the first incomplete vocal fold closure during the offset period ($X_{3.5g}$) for males at the mean and median times of the offset of the acoustic signal were 306.33ms and 312.5ms, respectively. For females we estimate the mean and median of ($X_{3.5g}$) as 312.39ms and 303.38ms respectively. Lastly, the estimated timing of the complete termination of vocal fold oscillation (X_{4g}) at the mean and median of the offset of the acoustic signal for males was found to be 327.9ms and 333.55ms, respectively. For females we estimate the mean and median of (X_{4g}) as 328.96ms and 333.55ms respectively. Thus during the offset period we find that on average the first instant of oscillatory offset (X_{3g}) occurs before the offset of the acoustic signal (X_{2a}) is detected ($p = 0.012$); the timing of the first incomplete vocal fold closure during the offset period ($X_{3.5g}$) occurs around the same time as the offset of the acoustic waveform for both genders ($p = 0.012$), with males slightly

earlier than females on average; and the complete termination of oscillatory motion (X_{4g}) occurs after the offset of the acoustic signal (X_{2a}) ($p = 0.012$) (Figure 3). Statistically significant gender differences were not observed across any offset variables, suggesting that gender does not predict the first instant of oscillatory offset (X_{3g}) ($p = 1.000$), the timing of the first incomplete vocal fold closure during the offset period ($X_{3.5g}$) ($p = 1.000$), nor the complete termination of oscillatory motion (X_{4g}) ($p = 0.456$) (Table 2).

Discussion

The goal of the current study is to establish the relationship between the onset/offset of the acoustic signal and various pre-phonatory/post-phonatory oscillatory events. The predicted relations between pre-phonatory vocal fold oscillations and acoustic onset as well as the expected timing between the offset of the acoustic signal and the end of oscillations were confirmed. Further, the examination of high-speed video images in relation to acoustic events is in some ways consistent with EGG measures of VAT^{19,20,34} and VRT.²¹ These similarities as well as some differences are discussed below.

Relationship between onset of the acoustic signal and pre-phonatory oscillatory events

On average, the onset of the acoustic signal (X_{1a}) occurs 17 ± 8.75 milliseconds after the onset of the first oscillatory motion of the GAW (X_{1g}) and 30.9 ± 18.08 milliseconds prior to the onset of sustained oscillations (X_{2g}). The acoustic signal is most closely tied to the first contact of the vocal folds ($X_{1.5g}$) in that the acoustic signal starts, on average, 5.78 ± 6.57 milliseconds before the first contact of the vocal folds ($X_{1.5g}$), with this interval for females being an average of 5.22 milliseconds longer compared to males. To the best of our knowledge these are the first findings from a large-scale study of the relationship between the onset of the acoustic signal and the discrete pre-phonatory events in the GAW. Most large-scale studies of vocal fold contact ($X_{1.5g}$) used EGG to monitor the time lag between the first vocal fold contact on EGG signal (VAT) and onset of the acoustic voice signal.^{19,34} Results from those studies indicate VATs between 1–3 ms with some differences between men and women and between hard and breathy onsets.^{19,34} EGG is sensitive to changes in vocal fold approximation and, therefore, provides information about the timing of glottal opening and closing, however, pre-phonatory oscillatory patterns are generally not well captured by this procedure. It is also difficult to obtain the exact time of vocal fold closure on EGG.³⁵ Although previous studies have examined VAT measures from high-speed imaging, those studies included small samples of speakers and a variety of speech material, which limited the power and interpretation of those data.^{12,20}

The current investigation included 21 men and 35 women and found similar, but on average longer latencies between the acoustic signal and first vocal fold contact than are seen in EGG analysis. The finer grained temporal analysis in this study provided more information about when vocal fold oscillation generates an acoustic signal. Specifically, the findings reveal a strong trend that the acoustic signal's onset occurs prior to the first instant of vocal fold contact ($X_{1.5g}$) for both men and women although there is a significantly longer delay between the onset of the acoustic signal and the vocal fold contact for women than for men.

Oscillations prior to vocal fold contact probably cannot be extracted from EGG, a procedure that captures varying impedance associated with the amount of vocal fold contact.

There are some similarities and differences in the current results compared to previous investigations using the VAT. For example, Roark et al. (2012)¹⁹ measured VAT of about 2 ms for men, a value that is consistent with the present finding. Like the current study, Roark et al (2012)¹⁹ also found significant differences between VAT in men and women, however, the women in our study had VATs that were on an average estimated to be 5.22 ms longer than those of men whereas Roark et al.(2012)¹⁹ report female VATs as being about 1 ms shorter than for males. Watson et al. (2016)³⁴ did not find a significant gender effect during the production of the word “hallways,” although there was a trend toward longer VATs for women than for men. Some of the differences between the current study and findings from previous investigations may relate to instrumental differences (high-speed videoendoscopy versus EGG) in data acquisition or to the general variability of measures reflecting first vocal fold contact. The variability between the findings from the current study and the previous investigations may relate to differences in the tasks and individual differences in the timing between the laryngeal articulation relative to the supraglottal articulation. The task in the current study involved the production of the /h/ following an /i/ whereas studies by Roark et al (2012)¹⁹ were performed on sustained production of the vowel /a/. The vowel environment and loudness levels have been shown to have a speaker-specific effect on the voice onset and offset measures.⁴ It is a question for future studies, if the temporal relationship between the acoustic onset and the various pre-phonatory events are invariant across various vowel environments, type of speech materials, duration of the utterance, loudness differences, and most importantly, how vocal disorders affect the relationship between acoustic and phonatory onsets. Even though all subjects were phonating at typical loudness level, future studies should include more stringent control for loudness levels. Additionally, our findings reveal large intersubject variability in the instance of first vocal fold contact and the acoustic signal’s onset. Despite these minor differences between studies, it is important to note that acoustic onset is prior to the first vocal fold contact as is evident from the findings of the current study. Given the nature of the current experiment, it is beyond the scope of this current manuscript to know the physiological basis for the gender differences in the onset phase. Possible explanations for this difference could be due to increased prevalence of a posterior phonatory gap^{36,37} and small size of the laryngeal structures in females.^{38–40}

Clearly, the present study is only a first step in assessing the complex relationship between the glottal wave and the ensuing acoustic signal. Many additional experiments on healthy and voice-disordered speakers are needed before the acoustic signal can serve as an independent metric of vocal control. The equations that are used to model this relationship were fit to the present data; therefore, further data need to be collected to determine if the derived equations can be generalized to other speakers with normal voice. In addition, data from disordered speakers are needed to determine if glottal kinematics can be aligned with the acoustic signal. As previous research indicates,^{41,42} voice disorders may result in aperiodic signals that would make it difficult to determine voicing onset from the acoustic wave. These obstacles need to be evaluated and overcome to be able to apply the current data in clinical practice.

Despite these limitations, one of the goals of the current study and previous investigations of VAT^{19,20,34} is to serve as reference for evaluations of voice disorders. Although there are some differences between studies in the relation of the acoustic signal with reference to the first contact of the vocal folds, these variations are the order of a few milliseconds whereas it is expected that people with voice disorders will demonstrate much larger differences in this measure. Future studies that compare the current findings with those from people with voice disorders can test this prediction.

Relationship between offset of the acoustic signal and post-phonatory oscillatory events

In the current study, measures were made of the relationship between the acoustic and the oscillatory events at voicing offset by relating the acoustic output to the end of steady state oscillatory motion (X_{2g}); the time of incomplete glottal closure ($X_{2.5g}$) and the end of detectable vocal fold movement (X_{3g}). The study findings reveal that offset of the acoustic signal (X_{2a}) occurs 63 ± 30.18 ms after the end of steady state oscillations (X_{3g}), after 3.66 ± 15.61 ms after incomplete glottal closure ($X_{3.5g}$) and 36.89 ± 20.04 ms prior to complete cessation of oscillatory motion (X_{4g}).

As the onset of the acoustic signal was tied to vocal fold contact, the offset of the acoustic signal was dictated by the loss of such contact between the folds. Taken together, the timing of vocal fold contact and the acoustic onset and incomplete glottal closure and vocal fold offset suggest a tight coupling between glottal closure and generation of an acoustic signal. Both acoustic onset and acoustic offset of voicing are within a few milliseconds of the onset and offset of vocal fold contact. As with the relationship between X_{1a} and $X_{1.5g}$, the results for timing between the acoustic offset (X_{2a}) and the loss of vocal fold contact ($X_{2.5g}$) from the current study are somewhat different from Watson et al (2012)²¹ report of vocal release time. Watson et al (2012)²¹ noted that the EGG signal decayed about 20 ms prior to the end of the acoustic signal whereas there was a tighter coupling between the end of vocal fold contact as measured from the GAW and offset of the acoustic signal in the current study. In the current study, an average of 1 to 6 ms separated the release of vocal fold contact and the end of the acoustic pressure wave for men and women, respectively. A number of factors may be responsible for these differences. First, high-speed video imaging, as used in the current investigation may yield better resolution such that more precise coupling between the GAW and the acoustic signal can be determined. A second difference between the current study and that of Watson and his colleagues (2012)²¹ was in the speech sample used to measure vocal release time. Watson et al (2012)²¹ measured VRT from a sustained production of /a/ whereas participants in the current study produced a sequence of three repetitions of the syllable /hi/ with measurements taken from only the middle repetition of the sequence. These different speech samples may require different control on VRT in that the termination of vocal fold oscillations in a single sustained vowel may be unconstrained, as noted by Watson et al. (2012)²¹ whereas a sequence of /hi/ syllables would necessitate an abrupt release of vocal fold contact to produce the /h/ in the next syllable in the sequence. A final factor that needs to be considered is that timing between the acoustic offset (X_{2a}) and the loss of vocal fold contact ($X_{2.5g}$) may be variable, as noted for the timing between the onset of the acoustic signal (X_{1a}) and the first instant of vocal fold contact ($X_{1.5g}$), and the differences between studies may reflect this variability.

In contrast to the relationship between the X_{1a} and $X_{1.5g}$, there were no significant differences between men and women in the current study for any of the offset measures although the acoustic signal ended approximately 1 ms prior to the loss of vocal fold contact for men and about 6 ms after the end of vocal fold contact for women. Again, this can relate to the relative variability of these measures across individuals and speech samples.

Much of the limited research on voice offset is centered on using these indices for evaluation of pathological voices. The relatively large sample size in this study, coupled with relatively consistent results across very different methodologies suggest that there is a strong normative database to which disordered speakers can be compared. In general, voice offset is constrained by linguistic specifications for phoneme voicing or devoicing and suprasegmental variables. As such, investigations of the relationship between the X_{2a} and $X_{2.5g}$ may be more useful in these types of environments rather than in the unconstrained contexts used in this and previous studies.

Conclusion

The relationship between the start of the acoustic signal and first contact of the vocal folds varied between males and females. Gender differences were not observed between the start of the acoustic signal and the start of the first vocal fold oscillation and the start of steady-state phonation. For offset, gender differences were not observed between the offset of the acoustic signal and the following oscillatory events: end of steady state, first instance of incomplete glottal closure, and end of oscillations, suggesting that offset and onset are governed by different physiological factors. The emerging normative references regarding the timing of the onset and offset from this study may serve as a foundation to study these relationships in non-normal quality. The findings from this study could be used as reference values for easy/simultaneous onset, which is thought to be a balanced onset and thereby provide a framework for future investigations aimed at classifying different types of voice qualities (hard, breathy, soft).⁴³

Acknowledgments

The authors thank Reuben Walker for his assistance with data processing.

Sources of financial support or funding: This research was supported by NIH/NIDCD R03DC11360.

References

1. Liskar L, Abramson AS. A cross-language study of voicing in initial stops: Acoustical measurements. *Word*. 1964; 20:527–565.
2. Robb MP, Chen Y. Is /h/ phonetically neutral? *Clin Linguist Phon*. 2009; 23:842–855. [PubMed: 19891524]
3. Koenig LL. Laryngeal factors in voiceless consonant production in men, women, and 5-year-olds. *J Speech Lang Hear Res*. 2000; 43:1211–1228. [PubMed: 11063242]
4. Koenig LL, Mencl WE, Lucero JC. Multidimensional analyses of voicing offsets and onsets in female speakers. *J Acoust Soc Am*. 2005; 118:2535–2550. [PubMed: 16266174]
5. Lucero JC. A theoretical study of the hysteresis phenomenon at vocal fold oscillation onset-offset. *J Acoust Soc Am*. 1999; 105:423–431. [PubMed: 9921668]

6. Regner MF, Tao C, Zhuang P, Jiang JJ. Onset and offset phonation threshold flow in excised canine larynges. *Laryngoscope*. 2008; 118:1313–1317. [PubMed: 18401267]
7. Koster O, Marx B, Gemmar P, Hess MM, Kunzel HJ. Qualitative and quantitative analysis of voice onset by means of a multidimensional voice analysis system (MVAS) using high-speed imaging. *J Voice*. 1999; 13:355–374. [PubMed: 10498052]
8. Kunduk, Yan Y, McWhorter AJ, Bless D. Investigation of voice initiation and voice offset characteristics with high-speed digital imaging. *Logoped Phoniatr Vocol*. 2006; 31:139–144. [PubMed: 16966156]
9. Patel, Pickering J, Stemple J, Donohue KD. A Case Report in Changes in Phonatory Physiology Following Voice Therapy: Application of High-Speed Imaging. *J Voice*. 2012
10. Schutte, HK., Švec, JG., Sram, F. Videokymography. Imaging and quantification of regular and irregular vocal fold vibrations. In: McCafferty, G.Coman, W., Carroll, R., editors. *Proceedings of the XVI World Congress of Otorhinolaryngology, Head and Neck Surgery*. Sydney, Australia: Monduzzi Editore S.p.A. Bologna, Italy; 1997. p. 1739-1742.
11. Hong KH, Kim HK, Niimi S. Laryngeal gestures during stop production using high-speed digital images. *J Voice*. 2002; 16:207–214. [PubMed: 12150373]
12. McDonnell M, Sundberg J, Westerlund J, Lindestad PA, Larsson H. Vocal fold vibration and phonation start in aspirated, unaspirated, and staccato onset. *J Voice*. 2011; 25:526–531. [PubMed: 20951547]
13. Larsson H, Hertegard S, Lindestad PA, Hammarberg B. Vocal fold vibrations: high-speed imaging, kymography, and acoustic analysis: a preliminary report. *Laryngoscope*. 2000; 110:2117–2122. [PubMed: 11129033]
14. Granqvist S, Hertegard S, Larsson H, Sundberg J. Simultaneous analysis of vocal fold vibration and transglottal airflow: exploring a new experimental setup. *J Voice*. 2003; 17:319–330. [PubMed: 14513955]
15. Freeman E, Woo P, Saxman JH, Murry T. A comparison of sung and spoken phonation onset gestures using high-speed digital imaging. *J Voice*. 2012; 26:226–238. [PubMed: 21256709]
16. Mergell P, Herzel H, Wittenberg T, Tigges M, Eysholdt U. Phonation onset: vocal fold modeling and high-speed glottography. *J Acoust Soc Am*. 1998; 104:464–470. [PubMed: 9670538]
17. Braunschweig T, Flaschka J, Schelhom-Neise P, Döllinger M. High-speed video analysis of the phonation onset, with an application to the diagnosis of functional dysphonias. *Medical Engineering & Physics*. 2008; 30:59–66. [PubMed: 17317268]
18. Moore P. Motion picture studies of the vocal folds and vocal attack. *Journal of Speech Disorders*. 1939; 3:235–238.
19. Roark RM, Watson BC, Baken RJ, Brown DJ, Thomas JM. Measures of vocal attack time for healthy young adults. *J Voice*. 2012; 26:12–17. [PubMed: 21524562]
20. Orlikoff RF, Deliyski DD, Baken RJ, Watson BC. Validation of a glottographic measure of vocal attack. *J Voice*. 2009; 23:164–168. [PubMed: 18083343]
21. Watson BC, Roark RM, Baken RJ. Vocal release time: a quantification of vocal offset. *J Voice*. 2012; 26:682–687. [PubMed: 22480753]
22. Hirano, M. *Clinical examination of voice*. New-York: Springer-Verlag; 1981.
23. Patel R, Donohue KD, Unnikrishnan H, Kryscio RJ. Kinematic measurements of the vocal-fold displacement waveform in typical children and adult populations: quantification of high-speed endoscopic videos. *J Speech Lang Hear Res*. 2015; 58:227–240. [PubMed: 25652615]
24. Lohscheller J, Toy H, Rosanowski F, Eysholdt U, Dollinger M. Clinically evaluated procedure for the reconstruction of vocal fold vibrations from endoscopic digital high-speed videos. *Med Image Anal*. 2007; 11:400–413. [PubMed: 17544839]
25. Boersma P. Praat, a system for doing phonetics by computer. *Glott International*. 2001; 5:341–345.
26. Boll SF. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*. 1979; 27:113–120.
27. Cooke A, Ludlow CL, Hallett N, Selbie WS. Characteristics of vocal fold adduction related to voice onset. *J Voice*. 1997; 11:12–22. [PubMed: 9075172]

28. Roark RM, Watson BC, Baken RJ. A figure of merit for vocal attack time measurement. *J Voice*. 2012; 26:8–11. [PubMed: 21524561]
29. Patel, Walker R, Sivasankar M. Spatiotemporal quantification of vocal fold vibration following exposure to superficial laryngeal dehydration: A preliminary study. *Journal of voice*. In press.
30. Patel R. Vibratory onset and offset time in children: A laryngeal imaging study. *Int J Pediatr Otorhi*. in press.
31. Kunduk M, Yan Y, McWhorter AJ, Bless D. Investigation of voice initiation and voice offset characteristics with high-speed digital imaging. *Logoped Phoniatr Vocol*. 2006; 31:139–144. [PubMed: 16966156]
32. Bless DM, Hirano M, Feder RJ. Videostroboscopic evaluation of the larynx. *Ear Nose Throat J*. 1987; 66:289–296. [PubMed: 3622324]
33. Patel, Dailey S, Bless D. Comparison of high-speed digital imaging with stroboscopy for laryngeal imaging of glottal disorders. *Ann Otol Rhinol Laryngol*. 2008; 117:413–424. [PubMed: 18646437]
34. Watson BC, Baken RJ, Roark RM. Effect of Voice Onset Type on Vocal Attack Time. *J Voice*. 2016; 30:11–14. [PubMed: 25795369]
35. Colton RH, Conture EG. Problems and Pitfalls of Electroglottography. *Journal of Voice*. 1990; 4:10–24.
36. Patel, Dixon A, Richmond A, Donohue KD. Pediatric high speed digital imaging of vocal fold vibration: A normative pilot study of glottal closure and phase closure characteristics. *Int J Pediatr Otorhinolaryngol*. 2012; 76:954–959. [PubMed: 22445799]
37. Linville SE. Glottal gap configurations in two age groups of women. *J Speech Hear Res*. 1992; 35:1209–1215. [PubMed: 1494266]
38. Titze IR. Physiologic and acoustic differences between male and female voices. *J Acoust Soc Am*. 1989; 85:1699–1707. [PubMed: 2708686]
39. Kahane JC. A morphological study of the human prepubertal and pubertal larynx. *Am J Anat*. 1978; 151:11–19. [PubMed: 623031]
40. Hirano, M., Kurita, S., Nakashima, T. *Growth, Development, and Aging of Human Vocal Fold*. San Diego, California: College-Hill Press; 1983.
41. Cannito MP, Buder EH, Chorna LB, Dressler R. Acoustic measures of phonation during connected speech in adductor spasmodic dysphonia. *Otolaryngology*. 2012:1–8.
42. Jiang JJ, Zhang Y, MacCallum J, Sprecher A, Zhou L. Objective acoustic analysis of pathological voices from patients with vocal nodules and polyps. *Folia Phoniatr Logop*. 2009; 61:342–349. [PubMed: 19864916]
43. Koike Y, Hirano M, Von Leden H. Vocal initiation: acoustic and aerodynamic investigations of normal subjects. *Folia Phoniatr*. 1967; 19:173–182.

Figure 1.

Sounds pressure and glottal area waveform during the production of /hi/ in a 32 year old male. The points X_{1a} and X_{2a} represent the onset and offset of the acoustic signal respectively. X_{1g} , $X_{1.5g}$, X_{2g} represent the instances of onset of first oscillatory motion, first medial contact of the vocal folds, and onset of sustained oscillations respectively. X_{3g} , $X_{3.5g}$, and X_{4g} represent the first instance of oscillatory offset, first incomplete vocal fold closure, and complete cessation of oscillatory motion.

Figure 2.

Variable kernel density estimates and scatterplots with regression lines by gender. The columns represent the x-axis while the rows represent the y-axis. Both columns and rows of the figure present the data for both the vocal fold vibratory (X_{1g} , $X_{1.5g}$, X_{2g} , X_{3g} , $X_{3.5g}$, and X_{4g}) and auditory events (X_{1a} , X_{2a}). The points X_{1a} and X_{2a} represent the onset and offset of the acoustic signal respectively. X_{1g} , $X_{1.5g}$, X_{2g} represent the instances of onset of first oscillatory motion, first medial contact of the vocal folds, and onset of sustained oscillations respectively. X_{3g} , $X_{3.5g}$, and X_{4g} represent the first instance of oscillatory offset, first incomplete vocal fold closure, and complete cessation of oscillatory motion. Along the diagonal are the kernel density estimates of each variable, which provides a nonparametric representation of each variable's distribution. For each, this provides the shape of the data and offers a visualization of where along the x-axis the data are most representative.

Table 1

Correlations between oscillatory and acoustic onset / offset variables (N = 56)

	1	2	3	4	5	6	7	8
1 Onset of first oscillatory motion								
2 First contact of the vocal folds	0.99							
3 Start of sustained oscillations	0.97	0.98						
4 First instant of oscillatory offset	0.82	0.92	0.81					
5 First instant of vocal folds losing contact	0.85	0.85	0.84	0.95				
6 Complete termination of oscillatory motion	0.84	0.84	0.83	0.94	0.99			
7 Start of acoustic waveform	0.99	0.99	0.97	0.83	0.85	0.85		
8 Offset of acoustic waveform	0.86	0.86	0.84	0.95	0.99	0.98	0.86	

Table 2

Summary of multivariate regression analysis for variables predicting oscillatory onset and offset events of vocal folds (N = 56). The points X_{1a} and X_{2a} represent the onset and offset of the acoustic signal respectively. X_{1g} , $X_{1.5g}$, X_{2g} represent the instances of onset of first oscillatory motion, first medial contact of the vocal folds, and onset of sustained oscillations respectively. X_{3g} , $X_{3.5g}$ and X_{4g} represent the first instance of oscillatory offset, first incomplete vocal fold closure, and complete cessation of oscillatory motion.

Variable	X_{1g}			$X_{1.5g}$			X_{2g}			X_{3g}			$X_{3.5g}$			X_{4g}		
	B	SE	β	B	SE	β	B	SE	β	B	SE	β	B	SE	β	B	SE	β
X_{1a}	1.01*	0.02	0.98	1.00*	0.02	0.99	1.07*	0.06	0.95	-	-	-	-	-	-	-	-	-
Female	3.67	2.24	0.033	5.22*	1.63	0.05	10.26	4.65	0.08	12.63	7.90	0.08	-7.21	4.20	0.04	-11.11	5.36	0.07
X_{2a}	-	-	-	-	-	-	-	-	-	0.94*	0.05	0.93	1.01*	0.03	0.99	0.998*	0.04	0.97
R^2	0.98			0.99			0.92			0.87			0.97			0.94		

Note: β represents the standardized coefficients, B represents the coefficients, and SE represents the standard error.

* $p < 0.05$.

Table 3

Summary statistics of the onset and offset events. The points X_{1a} and X_{2a} represent the onset and offset of the acoustic signal respectively. X_{1g} , $X_{1.5g}$, X_{2g} represent the instances of onset of first oscillatory motion, first medial contact of the vocal folds, and onset of sustained oscillations respectively. X_{3g} , $X_{3.5g}$ and X_{4g} represent the first instance of oscillatory offset, first incomplete vocal fold closure, and complete cessation of oscillatory motion. All temporal measures are relative to the beginning of the analysis file and are reported in milliseconds.

	Female (n = 35)			Male (n = 21)		
	Mean (SD)	Median	95% Confidence Interval	Mean (SD)	Median	95% Confidence Interval
X_{1a}	89.96 (59.22)	69.87	69.62, 110.30	79.13 (40.83)	67.8	60.54, 97.71
X_{1g}	74.28 (61.57)	50.5	53.13, 95.43	59.63 (39.81)	45	41.51, 77.75
$X_{1.5g}$	97.66 (58.89)	79.75	77.44, 117.89	81.70 (41.62)	66.75	62.76, 100.65
X_{2g}	125.00 (62.85)	102.5	103.41, 146.59	103.13 (53.52)	85	78.77, 127.49
X_{2a}	318.67 (83.64)	309.7	289.94, 347.40	305.46 (73.16)	311.6	272.16, 338.77
X_{3g}	250.56 (80.98)	231	222.74, 278.37	250.79 (80.49)	262.7	214.14, 287.42
$X_{3.5g}$	312.33 (84.94)	302.75	283.15, 341.51	306.27 (77.23)	309.25	271.12, 341.43
X_{4g}	351.35 (84.90)	347.5	322.19, 380.52	349.36 (77.49)	351.5	314.09, 384.64