Accepted Manuscript

Analysis of phonation onsets in vowel production, using information from glottal area and flow estimate

Tiina Murtola, Jarmo Malinen, Ahmed Geneid, Paavo Alku

PII: S0167-6393(18)30296-6
DOI: https://doi.org/10.1016/j.specom.2019.03.007
Reference: SPECOM 2635

To appear in: *Speech Communication*

Received date: 31 August 2018
Revised date: 14 January 2019
Accepted date: 31 March 2019

Please cite this article as: Tiina Murtola, Jarmo Malinen, Ahmed Geneid, Paavo Alku, Analysis of phonation onsets in vowel production, using information from glottal area and flow estimate, *Speech Communication* (2019), doi: https://doi.org/10.1016/j.specom.2019.03.007

Analysis of phonation onsets in vowel production, using information from glottal area and flow estimate

Tiina Murtola[a)], Jarmo Malinen[b)], Ahmed Geneid[c)], and Paavo Alku[a)],

[a)]Department of Signal Processing and Acoustics, Aalto University, School of Electrical Engineering, P.O. Box 13000, FI-00076 Aalto, Finland
[b)] Department of Mathematics and Systems Analysis, Aalto University, School of Science, P.O. Box 11100, FI-00076 Aalto, Finland
[c)] Department of Otorhinolaryngology and Phoniatrics - Head and Neck Surgery, Helsinki University Hospital and University of Helsinki, Helsinki, Finland

Abstract

A multichannel dataset comprising high-speed videoendoscopy images, and electroglottography and free-field microphone signals, was used to investigate phonation onsets in vowel production. Use of the multichannel data enabled simultaneous analysis of the two main aspects of phonation, glottal area, extracted from the high-speed videoendoscopy images, and glottal flow, estimated from the microphone signal using glottal inverse filtering. Pulse-wise parameterization of the glottal area and glottal flow indicate that there is no single dominant way to initiate quasi-stable phonation. The trajectories of fundamental frequency and normalized amplitude quotient, extracted from glottal area and estimated flow, may differ markedly during onsets. The location and steepness of the amplitude envelopes of the two signals were observed to be closely related, and quantitative analysis supported the hypothesis that glottal area and flow do not carry essentially different amplitude information during vowel onsets. Linear models were used to predict the phonation onset times from the characteristics of the subsequent steady phonation. The phonation onset time of glottal area was found to have good predictability from a combination of the fundamental frequency and the normalized amplitude quotient of the glottal flow, as well as the gender of the speaker. For the phonation onset time of glottal flow, the best linear model was obtained using the fundamental frequency and the normalized amplitude quotient of the glottal flow as predictors.

Keywords: phonation onset; vowel production; high-speed videoendoscopy; glottal inverse filtering

1 Introduction

Voiced sounds are created by (quasi-)periodic vibration of the vocal folds, and they are a fundamental category of speech sounds in all spoken languages. Studying the production of these sounds is usually focused on steady vocal fold oscillations, and transient phenomena at onset and offset of vocal fold oscillations receive less attention, although understanding these phenomena has both clinical and technical relevance. This work makes use of a multichannel dataset of vowel production from healthy adults, comprising high-speed videoendoscopy (HSV) images, and electroglottography (EGG) and free-field microphone signals, to study the onset of vocal fold oscillation.

The dynamics of voice initiation is often characterized by estimating the time between the release of a stop consonant and start of voicing from an audio signal, i.e., through the concept of voice onset time (VOT). This measure is, however, a characteristic of the stop–vowel combination, and it is not applicable when voicing is initiated without a preceding vocal tract constriction. In the absence of the constriction, the onset of phonation can be characterized using the time required by the vocal fold oscillations to reach steady phonation. Direct observation of this rate of change can only be done using visual means, such as HSV and videokymography (Švec and Schutte, 1996). These visual means, HSV in particular, are irreplaceable in both clinical and research work, and several studies have used HSV to study the onset of phonation (e.g., Mergell *et al.*, 1998; Braunschweig *et al.*, 2008; Patel *et al.*, 2017a). However, the invasiveness of HSV and the expertise required of the experimenter impose restrictions for its use.

HSV images require processing before they can be used to study phonation onsets. Vocal fold displacement trajectories can be computed by tracking one or more points on the vocal folds to obtain digital kymograms (Mergell *et al.*, 1998; Braunschweig *et al.*, 2008; Patel *et al.*, 2017b). This approach can provide an accurate description of the movement of discrete points in the vocal folds, but it may miss, e.g., incomplete closure of the glottis. In contrast, glottal area waveforms (GAW) computed from HSV data (Petermann *et al.*, 2016; Patel *et al.*, 2017a,b) represent the whole two-dimensional projection of the orifice between the vocal folds but without any information about the location of the glottal gap in the anterior–posterior direction. A third option for investigating phonation onsets using HSV data utilizes several kymograms to estimate the vibrating length of the vocal folds (Ikuma *et al.*, 2016).

Vocal fold oscillation onsets have been estimated from HSV data (kymograms or GAWs) using peak detection or amplitude thresholding (e.g., Wittenberg *et al.*, 1997), thresholding of the oscillating length of the vocal folds (Ikuma *et al.*, 2016; Kunduk *et al.*, 2017), and amplitude envelope fitting (e.g., Mergell *et al.*, 1998; Braunschweig *et al.*, 2008). For the latter purpose, Mergell *et al.* (1998) derived an envelope function from a bifurcation model of the vocal fold dynamics, which they then fitted to HSV data. This Mergell envelope, and its rate of growth, typically quantified with *phonation onset time* (POT), are often treated as the baseline against which other onset measures are compared. While envelope functions can be fitted directly to vocal fold displacement or GAW peaks (Mergell *et al.*, 1998; Petermann *et al.*, 2016), amplitude envelopes, computed via Hilbert transform, have been used as an intermediate step in Braunschweig *et al.* (2008) and Patel *et al.* (2017b). These amplitude envelopes are called Hilbert envelopes (HEs). HEs are the magnitudes of analytic signals which have been obtained from the time-domain waveforms using Hilbert transform (see, e.g., Oppenheim and Schafer, 1989: Chap. 10, pp. 662–694).

Comparisons between different onset duration measures have been carried out in Petermann *et al.* (2016) and Patel *et al.* (2017b): Petermann *et al.* (2016) used GAWs extracted from HSV data, and they compared the performance of the Mergell envelope and polynomial envelopes with different degrees, as well as the impact of different preprocessing methods of the HSV data. In addition to POT and durations derived from polynomial envelopes, Patel *et al.* (2017b) also included a duration based on changes in the amplitude periodicity of the GAW.

As an alternative to HSV-based measures, Orlikoff *et al.* (2009) proposed the use of vocal attack time (VAT), which is the time between an increase in the sound pressure in an acoustic signal and the corresponding onset in an electroglottography (EGG) signal (see also Watson *et al.*, 2013, 2016). Although their results indicate a correspondence between VAT and manually extracted onset duration in HSV data, obtaining reliable EGG signal can be challenging. Patel *et al.* (2017a) compared three manually extracted time instants in HSV data (first detected oscillation of vocal folds, first medial vocal fold contact, and sustained phonation) to the first periodic deviation in the acoustic signal. Their results indicate a quantifiable relationship between onsets in HSV data and acoustic signals, but the manual extraction of the time instants is subject to human error and judgment as well as to noise.

Even though multichannel data has been used to study the onset of phonation (Orlikoff *et al.*, 2009; Patel *et al.*, 2017a), the acoustical excitation of voiced speech generated by the vibrating vocal folds interacting with fluid dynamic and acoustic phenomena, the glottal flow (i.e., the volume velocity waveform), has not been used in these investigations. Indeed, despite the fact that glottal flow is an essential part of phonation, providing a link between vocal fold vibrations and produced speech signals, only its low-frequency components have been studied at phonation onsets (Hammer, 2013). Further, Hammer (2013) used a stop–vowel combination; hence, their results are not comparable to the vowel onsets typically used in HSV studies. The absence of studies utilizing glottal flow at phonation onsets can be explained by the infeasibility of measuring it directly in practice. However, glottal inverse filtering (GIF) provides a tool that can be used to estimate the glottal flow from audio signals. Although GIF has been widely used to study different aspects in steady phonation in speech (e.g., Holmberg *et al.*, 1988; Childers and Ahn, 1995) and singing (e.g., Sundberg *et al.*, 2005), its use in studying phonation onset has not been previously reported. Therefore, the general goal of the present study is to further general understanding of onset phenomena in vowel production by simultaneously analyzing glottal area and flow estimate, the two interlinked but generally not identical components of phonation.

The approach taken in this study focuses on two general aspects of phonation onsets: increase in amplitudes and changes in glottal pulse shapes. The following aims were set to facilitate the investigation of these aspects. First, by using simultaneous multichannel recordings of vowel productions, the purpose of this study is to compare changes in glottal pulse shapes during phonation onsets *qualitatively* between GAWs (estimated from HSV data) and glottal flows (estimated with GIF from simultaneously recorded audio signals). Second, the study aims to develop a *quantitative relationship* between the key onset feature parameters, related to amplitude changes, extracted from glottal area and glottal flow. These quantitative comparisons serve to show to what extent the two signals provide independent information about onsets.

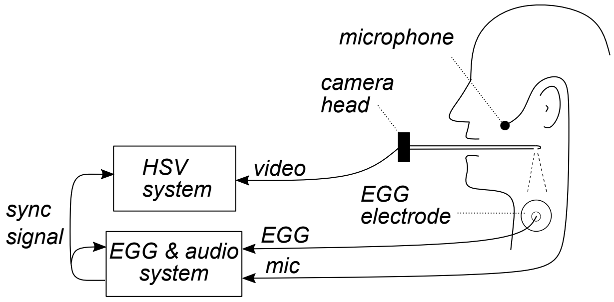


Figure 1: Vocal fold movements are recorded using a rigid endoscope connected to the HSV system. Simultaneous EGG and microphone signals are also acquired. A custom synchronization signal is recorded with the video, EGG, and microphone signals.

2 Data collection and processing

2.1 Data collection and exclusion

The data used in this investigation is a part of a larger, recently collected multichannel (HSV, audio, EGG) dataset which was originally designed for analysis of steady phonation. After the acquisition of this larger dataset, it was noticed that the data also included onsets in which the different data modalities behaved in a notably consistent manner. The larger dataset, described in more detail in Murtola *et al.* (2018), was designed as follows: Five male and five female speakers were instructed to vocalize a vowel sound using normal (i.e., modal) and breathy phonation at low, medium, and high pitch, in order to produce six perceptually different utterances (i.e., total of 60 utterances) at comfortable loudness. The production of the utterances was monitored, but speakers were free to choose comfortable pitch levels and degrees of breathiness. In order to obtain the clearest possible view of the glottis, the speakers were instructed to produce the Finnish vowel [i] with their tongues as far forward as possible. The HSV endoscope, however, hinders articulation which caused variance in the produced utterances so that they ranged between Finnish vowels [æ] and [œ]. Once phonation had been started, a pedal press by the experimenter triggered storing of the previous 4 s in the HSV system.

The setup for the data collection is shown schematically in Fig. 1. The HSV recordings were made using the KayPentax Color High-Speed Video System (model 9710) with a rigid endoscope. Spatial resolution of the video images was 512 x 512 pixels and temporal resolution 2000 frames/s. A Glottal Enterprises electroglottograph (EG2-PCX2) and a DPA omnidirectional headset microphone (model 4065-BL) were used to capture EGG and audio signals, respectively. The microphone was measured to lie approximately 6.5 cm from the center of the speaker's mouth as shown in Fig. 1. A MOTU UltraLite-mk3 Hybrid audio interface was used to record the microphone and EGG signals at sampling rate 44.1 kHz. The audio interface was connected to a MacBook Pro (OS X, v. 10.9.5), and AudioDesk 4 was used as the measurement software. A custom signal containing binary frequency-shift keyed code at the beginning of each second was used to synchronize the recordings. This signal was played during each measurement and recorded with both the HSV video and the audio–EGG signal pair.

High-pass filtering (cut-off frequency 60 Hz, linear phase finite impulse response (FIR) filter) was carried out on the audio and EGG signals. The data was synchronized by aligning the synchronization signals in the HSV data and the audio–EGG signal pair. The

Table 1: Data after selection. Sample ID is used to identify the samples used in this work. Speaker ID differentiates between the speakers, and these labels are the same as in the dataset of steady phonation.

Sample ID	Gender	Speaker ID	Pitch task	$\bar{f}_o$ (Hz)	Phonation task
m01	male	M01	low	110	breathy
m02	male	M01	medium	106	breathy
m03	male	M01	medium	122	normal
m04	male	M02	high	205	breathy
m05	male	M02	medium	101	breathy
m06	male	M02	low	95	normal
m07	male	M03	medium	141	breathy
m08	male	M05	medium	111	breathy
f01	female	F03	medium	229	normal
f02	female	F04	low	187	normal
f03	female	F04	medium	286	normal

latter were shifted to account for propagation delays (approximately 1.6 ms for males and 1.5 ms for females) and internal delays within and between the measurements systems. The maximum remaining error in the synchronization is ±0.5 ms (one frame in either direction) between the EGG signal and the video, and ±0.08 ms between the EGG and audio signals.

The data included a total of 13 onsets, and a frame of 200 ms surrounding each was analyzed. After exclusion of the data, where the vocal folds are not fully visible or the microphone signal was contaminated by external disturbances, 11 samples containing the onset of vocal fold oscillations remain (Table 1). For this work, each sample is considered to contain three conceptually different segments, which may or may not have transition regions between them: (i) *pre-phonation segment* has no clear periodic activity, (ii) *phonation initiation segment* is where periodic activity emerges and its amplitude increases rapidly, and (iii) *stabilization segment* contains slowly changing or stationary amplitudes and waveform shapes. The main focus of this investigation is on the phonation initiation segment, and a precise procedure to define this segment is detailed in Section 3.3. The pre-phonation and stabilization segments are named for ease of describing phenomena which are observed before or after the segment of interest, and hence their precise definitions are not needed. It is worth noting that phonation in the stabilization segment would generally be considered steady and, thus, suitable for conventional approaches to investigating vowel production.

2.2 Glottal area extraction

The GAW, $A(t)$, was extracted frame by frame from the red channel of the color video using the adapted seeded region growing method developed by Lohscheller *et al.* (2007). The extracted GAWs were manually inspected and, where necessary, corrected to counteract the inaccuracies introduced by light reflected from the closed glottis which caused periodical changes in illumination.

HSV and microphone recordings were carried out using different sampling frequencies (2 kHz and 44.1 kHz, respectively) so that resampling to a common timebase was required to carry out meaningful comparisons. GIF requires that the data be sampled at 8 kHz or higher (Alku and Vilkman, 1995); hence, the common sampling rate was selected to be 10 kHz. The GAWs were upsampled using MATLAB's function `resample` with default settings, i.e., an antialiasing low-pass FIR filter and delay compensation. This upsampling perserves the original frequency contents of the signal and introduces no temporal distortions (see, e.g., Oppenheim and Schafer, 1989: pp. 101–112). However, some fluctuations may be seen in the signal during the closed phase of the glottal cycle, and these were removed by forcing the resampled GAWs to be zero when the 2 kHz signals were zero, as well as anywhere where the resampled area signal was negative. The interpolated points in the GAWs have a larger margin of error than the measured points. During phonation onsets, when the glottis typically remains partially open and there are no abrupt changes in the pulse shapes (i.e., closures), the quality of the resampled signal is good throughout.

2.3 Glottal flow estimation using inverse filtering

The microphone and EGG signals were downsampled from their original sampling rate of 44.1 kHz to the selected common sample rate of 10 kHz using MATLAB's function `resample` with default settings. Inverse filtering of the microphone signal was carried out using Aalto Aparat (Alku *et al.*, 2017), which is a semi-automatic GIF tool. Aalto Aparat allows the key GIF parameters to be adjusted by the user in order to produce both the estimated glottal flow $U(t)$ and its first time derivative as time domain waveforms. Two GIF methods are available in Aalto Aparat: iterative adaptive inverse filtering (Alku, 1992) and quasi-closed phase (QCP) analysis (Airaksinen *et al.*, 2014). The latter was used in the current study because, when compared with four other common GIF algorithms, it was observed to be the most accurate in Airaksinen *et al.* (2014). The EGG signals were used to support GIF by visually checking that glottal openings and closures were aligned in $U(t)$ and the EGG signal. EGG was used for this purpose instead of HSV due to its smaller maximum synchronization error with the audio signal.

In order to obtain the glottal flow estimate, a frame containing the stabilization segment of each sample (as defined in Section 2.1) was selected manually in Aalto Aparat, and this frame was used to find the GIF parameters. These parameters were then used to obtain the glottal flow estimate for the entire sample. Although the pre-phonation segment affects estimation of the vocal tract filter model in this approach, the effect is negligible. This is due to the low amplitude level in the pre-phonation segment which causes the autocorrelation-based computation of the vocal tract model in QCP to focus automatically on the large-energy stabilization segment. Since the duration of the phonation initiation segment is short compared to exhalation time and time required for notable articulation, the vocal tract related GIF parameters extracted from the stabilization segment describe the phonation initiation segment as well.

3 Analysis methods: pulse-wise changes and amplitude envelopes

A common framework for the glottal area and glottal flow estimates at onsets of vowel production is formed by parameterizing both signals in terms of short-term *pulse-wise*

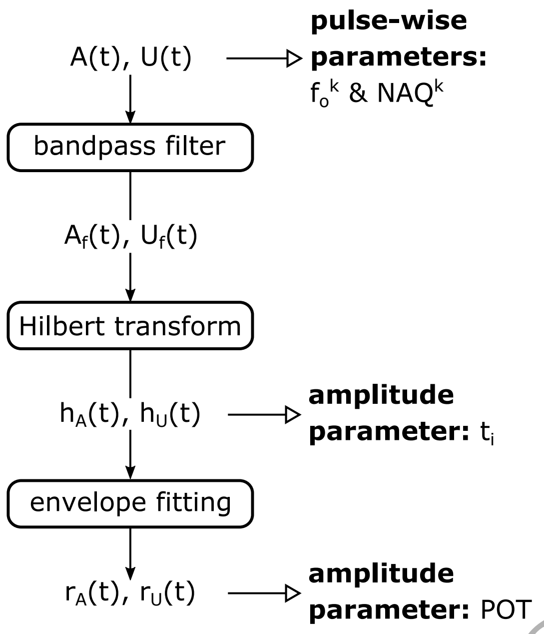


Figure 2: The main steps in extracting pulse-wise and amplitude envelope information from the glottal area $A(t)$ and glottal flow $U(t)$.

changes, i.e., changes from one glottal cycle to the next, and long-term *amplitude envelopes*. The main steps in the method are shown in Fig. 2. Since both the glottal area $A(t)$ and the glottal flow $U(t)$ are smooth and quasi-periodic time-domain waveforms, they will be treated equally as input $x(t)$ in the parameterization procedures below. Where necessary, subscripts are added to indicate whether a feature was extracted specifically from $A(t)$ or $U(t)$.

3.1 Pulse-wise parameters

Pulse-wise treatment of $x(t)$ was conducted using a procedure described below, which is similar to the algorithms in Aalto Aparat (for details, see Airas, 2008) but with some minor modifications to account for the transient nature of phonation onsets. Two pulse-wise parameters were computed for each identifiable pulse in the glottal area and flow signals. The first parameter, f_o^k, is a measure for the fundamental frequency of the k^{th} pulse in $x(t)$ ($k = 1, ..., N$, where N is the total number of pulses in a sample), and the second parameter, NAQ^k, is the normalized amplitude quotient (Alku *et al.*, 2002) which quantifies the shape of the k^{th} pulse.

Parameter f_o^k was computed from the time instants given by Aalto Aparat as

$$f_o^k = \frac{2}{(t_c^k - t_c^{k-1} + t_o^{k+1} - t_o^k)}, \quad k = 2, ..., N-1, \tag{1}$$

where t_c^k and t_o^k are the closing and opening instants in $x(t)$, respectively. When $k = 1$, only the opening instants were used, and when $k = N$, only the closing instants were used. Using the average of the fundamental period given by opening and closing instants makes f_o^k more robust against noise.

NAQ was selected as the pulse shape parameter for two reasons: First, it is a robust scalar quotient that has been shown in previous studies to be effective in parameterizing time-domain changes in the glottal flow when, for example, phonation type (Alku *et al.*, 2002), singing style (Björkner *et al.*, 2006), or vocal emotion (Airas and Alku, 2006) changes. Second, NAQ makes use of peak amplitude and the minimum of the derivative which can be identified in both glottal area and flow estimate using identical criteria. Although NAQ is conventionally used as a parameter for the shape of the glottal flow pulse, for this work, it is used to parameterize glottal area pulses as well. NAQ^k is computed as

$$\text{NAQ}^k = \frac{\max x^k(t) - \min x^k(t)}{|\min \dot{x}^k(t)|}\bar{f}_o, \tag{2}$$

where $x^k(t)$ is the waveform of the k^{th} pulse. Normalization of NAQ^k is done using average fundamental frequency $\bar{f}_o = \bar{f}_{o,U}$ which is computed by Aalto Aparat for the stabilization segment of $U(t)$ using the Yin method (de Cheveigné and Kawahara, 2002). In sufficiently long phonation, $\bar{f}_{o,U} \approx \bar{f}_{o,A}$, and this was checked to be true for the stabilization segments. The values of $\bar{f}_o$ are listed for each sample in Table 1.

3.2 Bandpass filtering and Hilbert transform

Vowel production is characterized by quasi-periodic $A(t)$ and $U(t)$, with a strong f_o component. In order to access amplitude information, which is mainly carried at a frequency component near f_o, $A(t)$ and $U(t)$ were bandpass filtered (linear phase FIR of order 200, cut-off frequencies $0.8\bar{f}_o$ and $1.2\bar{f}_o$, zero-phase filtering using MATLAB's `filtfilt`). The HSV and GIF methods used do not provide absolute amplitude values for the output signals; therefore both $A(t)$ and $U(t)$ were normalized to the range $[0, 1]$ after the bandpass filtering. It is worth noting, however, that all quantitative measures used in this work are scale invariant; hence, the scaling of the signals is only necessary for visual inspection of the data. The bandpass filtered and normalized versions of $A(t)$ and $U(t)$ are denoted $A_f(t)$ and $U_f(t)$, respectively.

The Hilbert transforms of $A_f(t)$ and $U_f(t)$ were obtained using the function `hilbert` in MATLAB with default settings. Amplitude envelopes were computed as the absolute value of the transform. The resulting HEs are denoted $h(t)$.

3.3 Envelope fitting and amplitude parameters

The mathematical function introduced by Mergell *et al.* (1998) was fitted to the HEs of $A_f(t)$ and $U_f(t)$ to obtain smooth parametric descriptions of the envelopes

$$r(t) = \pm r_0 \left([1 - \zeta]\, e^{-2at} + \zeta \right)^{-1/2}, \tag{3}$$

where $\zeta = r_0^2/r_\infty^2$, $r_0 = r(0)$, and $r_\infty = \lim_{t\to\infty} r(t)$. POT is defined using the parameter a: POT$= 1/a$, and it corresponds to amplitude growth of $r(t)$ from 32.2% to 67.8% (Mergell *et al.*, 1998).

Within each sample, the Mergell envelope $r(t)$ best describes the phonation initiation segment mentioned in Section 2.1. Therefore, $r(t)$ is only fitted to the part of the HEs corresponding to this segment which is identified through the derivative of the HE $\dot{h}(t)$. The inflection point $t_i = \arg\max \dot{h}(t)$ (see Figure 3) was first used to locate the onset in the signal. The phonation initiation segment was then defined to be the segment $[t_0, t_e]$ surrounding this point, where $\dot{h}(t) \geq 0.3 \max \dot{h}(t)$. This method was successful in

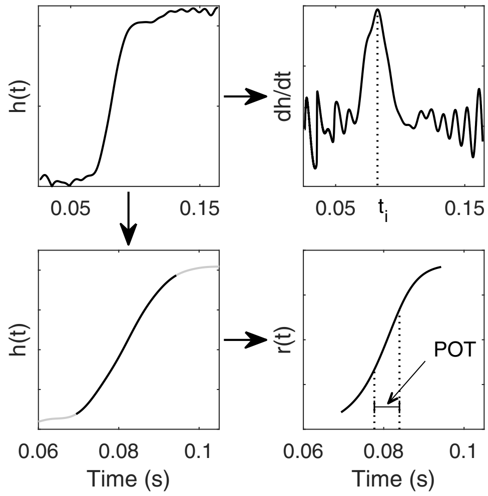


Figure 3: Extraction of amplitude parameters from HEs. Inflection point t_i is the instant of maximum time derivative of $h(t)$. POT is estimated by identifying the phonation initiation segment from $h(t)$ and fitting the Mergell envelope $r(t)$ to it.

identifying the segment with increasing amplitude envelope associated with the onset of phonation in all samples. However, in sample m05, automatic extraction of t_i may have placed it in a wrong location within the segment, as discussed later.

Optimization was carried out to minimize

$$f(t) = (h(t) - r(t - t_0))^2, \quad t \in [t_0, t_e] \tag{4}$$

using unconstrained optimization in MATLAB (function `fminunc` with default settings). All three parameters r_0, r_∞, and a in Eq. (3) were allowed to vary in the optimization.

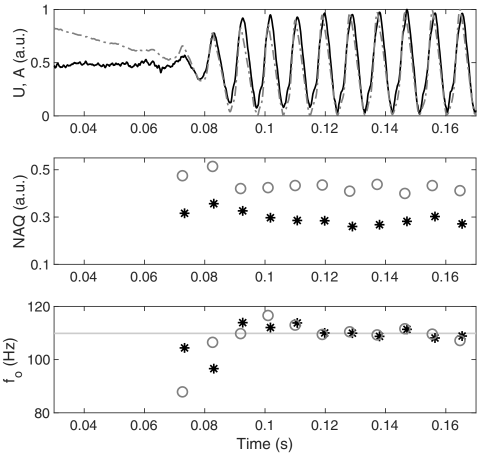
4 Results

Comparison of glottal flow and area at phonation onsets is presented below in two parts to match the two goals set at the end of Section 1 for this study. First, qualitative features are shown with a particular focus on pulse-wise characteristics. Second, quantitative comparisons of parameters related to the amplitude growth at onsets are proffered.

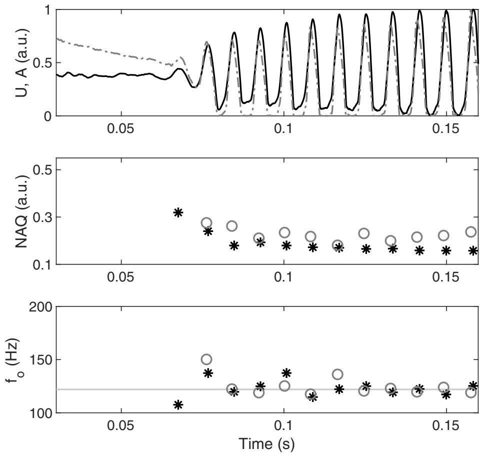
4.1 Onsets in glottal area and glottal flow

A selection of the glottal flows and GAWs are shown in Figs. 4—5. Before oscillations begin, some of the GAWs (topmost panels in Figs. 4 (a), (b), (d), and 5 (a)) display clear vocal fold abduction or adduction. These prephonatory gestures correspond mostly to silence in our data, with only sample m05 containing audible whispery sound before oscillations; hence, the glottal flow estimate is an abitrary constant during these gestures, i.e., GIF yields no information about the flow. However, once oscillations begin, the two waveforms become remarkably similar, especially at the beginning of the oscillations.

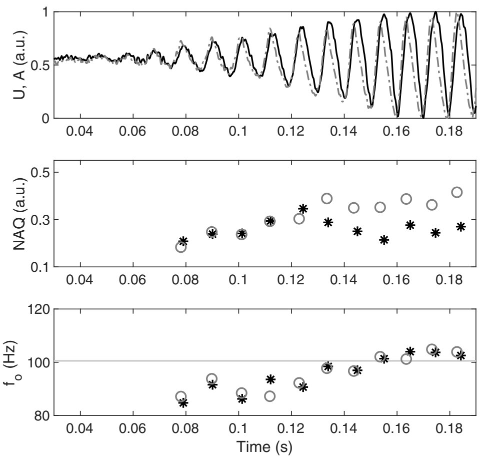
Figs. 4—5 also show pulse-wise parameters: normalized amplitude quotient NAQ^k and fundamental frequency f_o^k. Although NAQ values in steady phonation are indicative



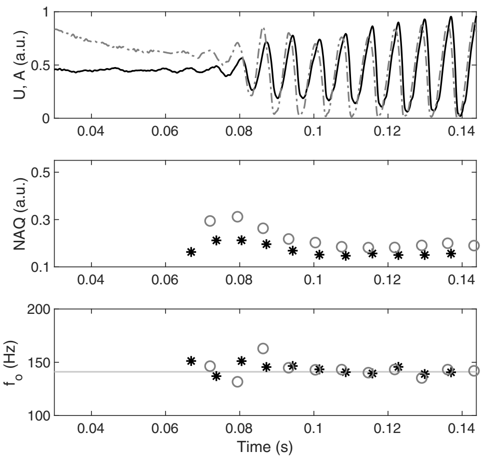
(a) m01: low pitch (110 Hz), breathy phonation



(b) m03: medium pitch (122 Hz), normal phonation

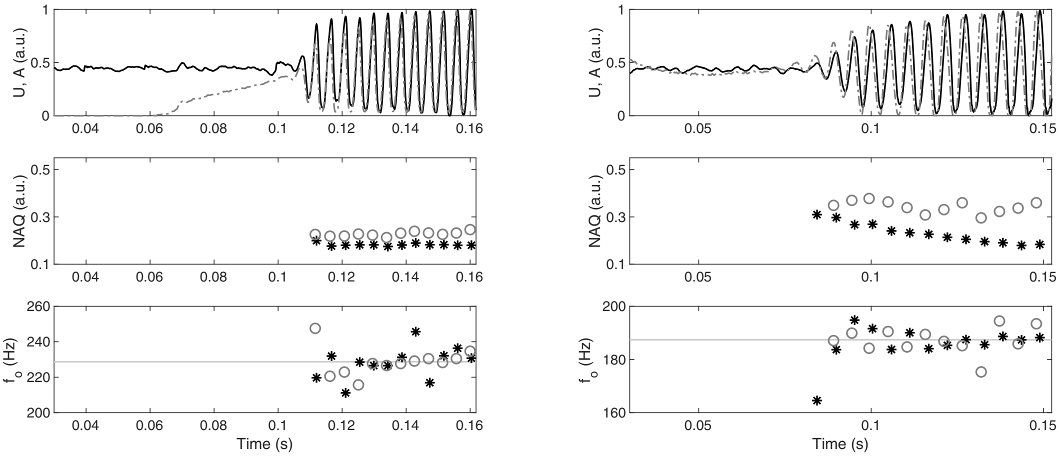


(c) m05: medium pitch (101 Hz), breathy phonation



(d) m07: medium pitch (141 Hz), breathy phonation

Figure 4: Phonation onsets in four representative samples by male speakers (a–d). For each sample, the top panel shows glottal flow U (solid black) and glottal area A (dashed gray); the middle panel shows NAQ^k extracted from U (black asterisk) and A (gray circles); the bottom panel shows f_o^k extracted from U (black asterisk) and A (gray circles), as well as the stabilized fundamental frequency $\bar{f}_o$ (horizontal line). For corresponding pitch and phonation mode tasks, see Table 1.



(a) f01: medium pitch (229 Hz), normal phonation

(b) f02: low pitch (187 Hz), normal phonation

Figure 5: Phonation onsets in two representative samples by female speakers (a–b). For each sample, the top panel shows glottal flow U (solid black) and glottal area A (dashed gray); the middle panel shows NAQ^k extracted from U (black asterisk) and A (gray circles); the bottom panel shows f_o^k extracted from U (black asterisk) and A (gray circles), as well as the stabilized fundamental frequency $\bar{f}_o$ (horizontal line). For corresponding pitch and phonation mode tasks, see Table 1.

of the mode of phonation, the rapidly changing amplitude of $A(t)$ and $U(t)$ within a single glottal cycle can dominate NAQ values at the beginning of phonation initiation (e.g., Fig. 4 (a), (c), and (d)). In most of the samples, there is a local maximum in the NAQ values during the phonation initiation indicating a soft closing phase. This occurs when the amplitude has increased, but the speed of closure is still relatively low. The decrease in NAQ observed after this maximum is due to faster closure.

There is also a clear tendency for NAQ$_A$ to be higher than NAQ$_U$, i.e., glottal flow pulses are more skewed to the right than the area pulses. Similar skewing of the glottal flow has also been observed in, e.g., Childers *et al.* (1985) and Hertegård and Gauffin (1995). In many, though not all, samples, this difference is more evident during the stabilization segment than during the phonation initiation segment. Fig. 6 illustrates this unequal skewing process through a Lissajous plot of sample m03: As oscillations begin, glottal flow and area are fairly close to the line $A(t) = U(t)$ but as phonation moves towards stabilization, the trajectory diverges increasingly from this line.

The pulse-wise f_o trajectories do not show a systematic pattern of reaching a stable level. There is, however, some indication that pulse-to-pulse changes in f_o, as well as the difference between the f_o^k values extracted from $A(t)$ and from $U(t)$, tend to be larger during initiation than late stabilization.

It is worth noting that some of the fluctuations seen in the pulse-wise parameter values, in particular f_o^k, in Figs. 4–5 may be caused by noise and estimation errors in the signals. The upsampled GAWs have their highest uncertainty at the points of glottal closure, which are used in computing f_o^k values. If the closing instants could be located only with the accuracy of one frame of HSV video (i.e., upsampling yielded no additional information), the error bounds for the f_o^k estimation would be approximately $\pm 0.05\bar{f}_o$

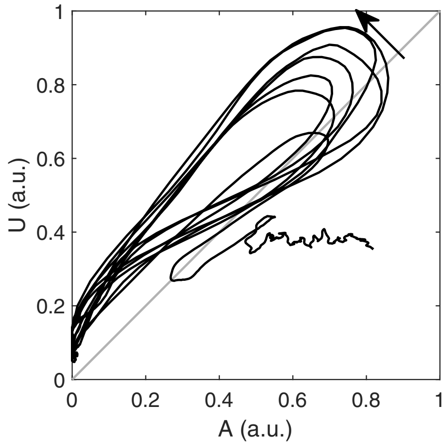


Figure 6: Lissajous plot of glottal flow versus glottal area for first several pulses of sample m03 (medium pitch (122 Hz), normal phonation). Arrow indicates direction of increasing time, and $U = A$ is shown as a diagonal line.

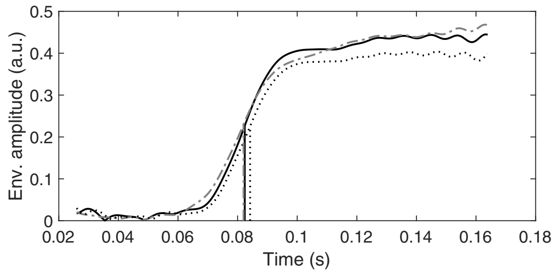
when $\bar{f}_o = 100$ Hz and $\pm 0.11\bar{f}_o$ when $\bar{f}_o = 200$ Hz. Similarly, the glottal flow estimates may have a formant ripple in closed phase, a known artifact of GIF (Alku, 2011) caused by imperfect cancellation of the vocal tract, which makes accurate estimation of opening and closing instants challenging. This effect tends to be more pronounced at high pitches as well, explaining the discrepancies between $f^k_{o,A}$ and $f^k_{o,U}$ in Fig. 5.

The NAQ values are less sentitive to the effects of the relatively low original HSV frame rate than f^k_o, as the vocal fold physiology favors low-frequency components in the oscillations. This is particularly true during early onset, as well as in breathy phonation, where the vocal folds do not close completely. When glottal closure occurs, the high uncertainty in the GAW at that instant may translate to uncertainty in the minimum derivative required for NAQ computation, particularly when f_o is high. However, abrupt changes in the NAQ values, caused by these errors when full glottal closure starts to occur during the onset, are not visible in the data.

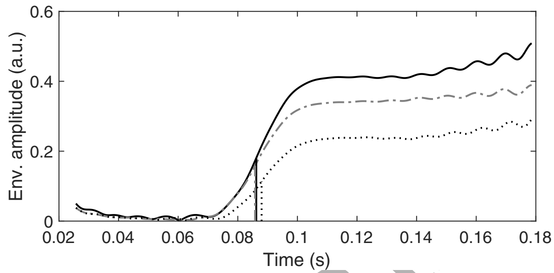
Pulse-wise parameters are, by their definitions, best suited to characterizing stable waveforms. As Figures 4–5 show, they can be used to parameterize phonation onsets, but interpretations of their values need to take into consideration the rapid amplitude growth occurring at the phonation onset.

4.2 Amplitude envelopes and POT

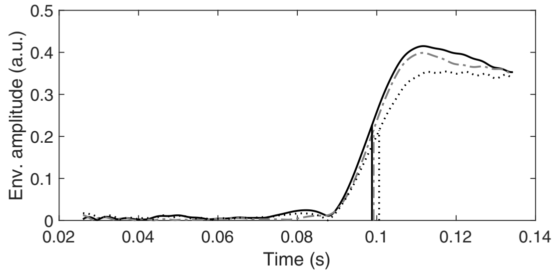
The HEs of $A(t)$ and $U(t)$ are shown in Fig. 7 for representative samples. The figure also shows the inflection instants $t_{i,A}$ and $t_{i,U}$ of $h_A(t)$ and $h_U(t)$, respectively. In addition, the HE of the audio signal h_M (i.e., without first estimating $U(t)$) and its inflection instant $t_{i,M}$ are also shown. The audio signal carries information about the vocal tract resonances, which is absent from the glottal signals, and therefore h_M scales differently than h_A and h_U. Since all three signals are bandpass filtered before computation of the HEs, the potential impact of the originally different sampling rates on amplitudes is removed. It is worth noting, however, that all measures used to characterize the onset in this work are scale invariant; hence, mismatches between the scales of the envelopes do not affect the numerical results.



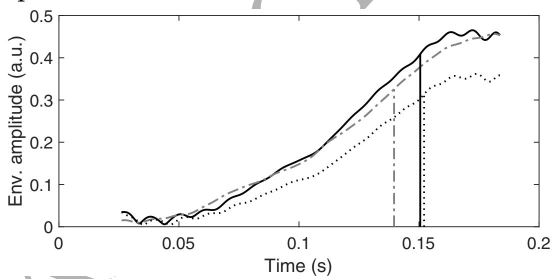
(a) m01: low pitch (110 Hz), breathy phonation



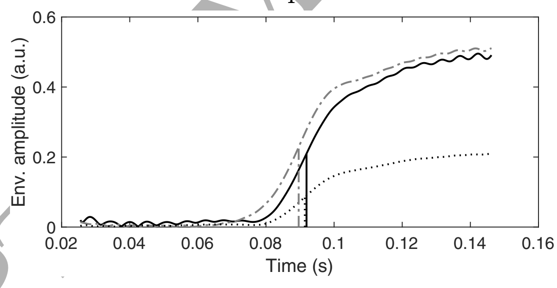
(b) m02: medium pitch (106 Hz), breathy phonation



(c) m04: high pitch (205 Hz), breathy phonation



(d) m05: medium pitch (101 Hz), breathy phonation



(e) f02: low pitch (187 Hz), normal phonation

Figure 7: Hilbert envelopes for four samples from male speakers (a–d) and one sample from a female speaker (e). For each sample, the envelopes extracted from U (solid black), A (dashed gray), and audio signal (corresponding to subscript M, dotted black) are shown. Vertical lines indicate the inflection point of each envelope.

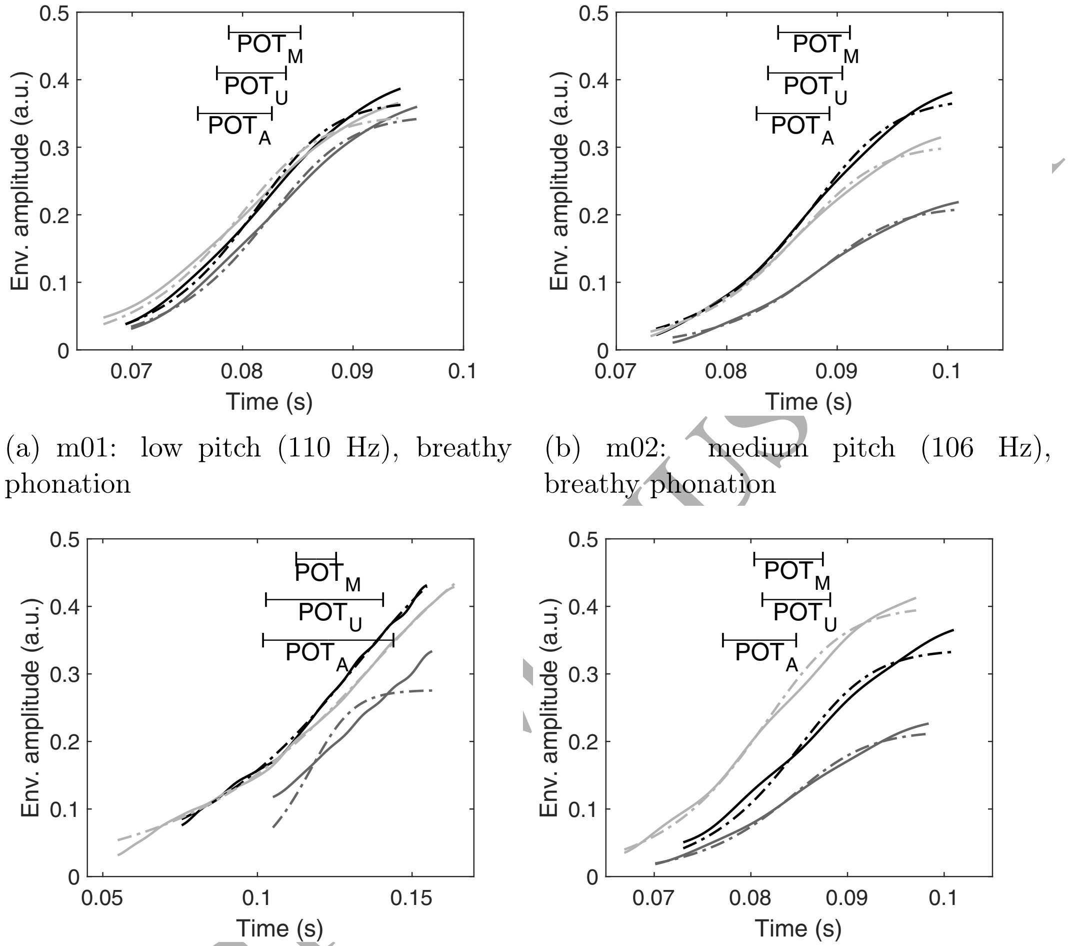
Table 2: Parameters of amplitude envelopes: differences between the inflection instant of the HEs and POT values of the fitted Mergell envelopes.

Sample ID	$t_{i,U} - t_{i,A}$ (ms)	$t_{i,M} - t_{i,A}$ (ms)	POT_A (ms)	POT_U (ms)	POT_M (ms)
m01	0.4	2.2	6.70	6.25	6.50
m02	0.4	2.2	6.60	6.73	6.50
m03	0.4	-0.8	4.80	5.03	5.30
m04	-0.4	1.4	4.58	4.75	5.23
m05	10.9	12.5	42.17	37.88	12.90
m06	8.8	5.3	7.58	7.03	7.10
m07	0.6	2.9	4.98	5.30	5.35
m08	1.0	1.1	6.48	6.88	7.53
f01	0.2	-1.7	4.78	3.98	3.95
f02	2.3	1.8	5.73	6.08	5.73
f03	0.8	1.6	5.80	6.18	5.38
Mean	2.31	2.59	9.11	8.73	6.50
SD	3.82	3.76	11.01	9.71	2.35
Mean excl. m05	1.45	1.60	5.80	5.82	5.86
SD excl. m05	2.67	1.92	1.02	1.01	1.05

The t_i values could be found automatically with no *a priori* information. However, m05 had several $\dot{h}(t)$ maxima of nearly equal magnitude; hence, the desired inflection point in m05 was not as clearly identifiable as in the other samples. The values of $t_{i,U}$ and $t_{i,M}$ relative to $t_{i,A}$ are listed in Table 2 together with the key statistics of each time difference. All information of interest is contained in these two time differences as the absolute location of the onsets within each sample is arbitrary.

In ten out of the eleven samples, $t_{i,A} < t_{i,U}$, whereas $t_{i,A} < t_{i,M}$ in nine of the samples. However, in five samples $|t_{i,A} - t_{i,U}| \leq 0.5$ ms, which is the maximum synchronization error between the signals, i.e., the difference may be caused by uncertainties in the synchronization. The lower temporal resolution of the HSV data is unlikely to be a major cause of error in the $t_{i,A}$ values, as both bandpass filtering and computation of HEs mitigate upsampling errors. In order to summarize the results on the two time differences quantitatively, one-sided paired sign tests were carried out with $\alpha = 0.05$. This nonparametric statistical test was chosen due to the small sample size ($N = 11$) and potential asymmetry of the differences. The tests indicate that both time differences, $t_{i,U} - t_{i,A}$ and $t_{i,M} - t_{i,A}$, are statistically significantly larger than zero ($p = 0.006$ and $p = 0.033$, respectively). Since the inflection points in m05 may have been misidentified, the tests were repeated with this sample excluded ($N = 10$). The value of $t_{i,U}$ remained significantly larger than $t_{i,A}$ ($p = 0.011$) but the difference between $t_{i,M}$ and $t_{i,A}$ became nonsignificant ($p = 0.055$). Overall, the results indicate that onsets in the glottal flow and the acoustic voice signal (as indicated by the time of maximum amplitude growth) tend to occur later or slower than vocal fold oscillation initiation, even after accounting for propagation delays. This difference is, however, typically only a couple of milliseconds.

The Mergell envelopes $r(t)$ from (3) fitted to the initiation segment of the HEs can be seen in Fig. 8. These figures also show the POT values computed from the Mergell envelopes. Since $r(t)$ is optimized to the initiation segment only, the fitting excludes the



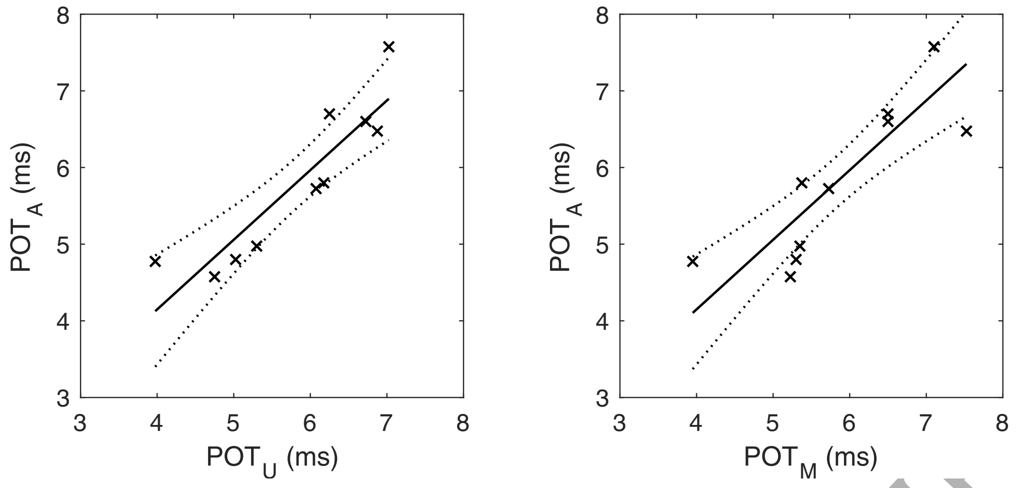
(a) m01: low pitch (110 Hz), breathy phonation

(b) m02: medium pitch (106 Hz), breathy phonation

(c) m05: medium pitch (101 Hz), breathy phonation

(d) m06: low pitch (95 Hz), normal phonation

Figure 8: Phonation initiation segments for four representative samples from male speakers (a–d). For each sample, Hilbert envelopes of U (solid black), A (solid light gray), and audio signal (corresponding to subscript M, solid dark gray) are shown, as well as the Mergell envelopes fitted to each Hilbert envelope (dashed lines). Horizontal bars indicate the phonation onset times extracted from the Mergell envelopes.



(a) (b)

Figure 9: Linear model for POT_A versus (a) POT_U and (b) POT_M. Solid line is the fitted model and dashed lines indicate 95% confidence bounds.

typically noisy transitions region from pre-phonation to initiation as well as the stabilization segment where the dynamics of saturation do not necessarily follow the form of $r(t)$. Despite this, the shape of $r(t)$ appears non-ideal for describing the HEs, particularly at the beginning and the end of the fitted segment. It is worth noting that while the Mergell envelope is able to match the uncommon shapes of $h_A(t)$ and $h_U(t)$ of m05, the optimization has failed in the case of $h_M(t)$ of the same sample.

The discrepancy between Mergell function and HE appears to be highly systematic; hence, comparison of the POT values computed from Mergell envelopes of $A(t)$ (POT_A), $U(t)$ (POT_U), and the audio signal (POT_M) is meaningful. These three POT values are listed in Table 2 and plotted in Fig. 9. Pearson correlation coefficients are high between POT_A and both POT_U and POT_M ($r = 0.999$ and $r = 0.933$, respectively); therefore, linear models were fitted between the POT values using linear regression. This yielded

$$\mathrm{POT}_A = 0.909 \cdot \mathrm{POT}_U + 0.515\,\mathrm{ms} \tag{5}$$

and

$$\mathrm{POT}_A = 0.813 \cdot \mathrm{POT}_M + 1.04\,\mathrm{ms}, \tag{6}$$

which predict POT_A from POT_U and POT_M with R^2 of 0.818 and 0.712, respectively, and maximum (absolute) residuals of 0.678 ms and 0.762 ms (Fig. 9). The data for m05 has been excluded from these models, as its POT_A and POT_U values are an order of magnitude larger than for the rest of the samples, and hence dominate least-squares models.

Even excluding m05, POT_A values range from 4.6 ms to 7.6 ms and POT_U from 4.0 ms to 7.0 ms. In order to investigate quantitative relationships between these POT values and target phonation, linear regression models were fitted between the POT values and parameters extracted from the stabilization segment. The resulting linear models with different predictor combinations have been compared in Table 3. $\overline{\mathrm{NAQ}}_A$ and $\overline{\mathrm{NAQ}}_U$ are seven-pulse averages taken from the end of the stabilization segment, and $\bar{T}_o = 1/\bar{f}_o$. The table shows only those predictor combinations that result in a model which has a significantly lower sum of squared errors than an intercept-only model (overall F-test for regression with $\alpha = 0.05$) for at least one of the POT values.

The fundamental period $\bar{T}_o$ is the single most powerful predictor of both POT values (Table 3) indicating that the majority of the amplitude growth for the used phonation task tends to happen in a constant number of glottal cycles. POT_A can be predicted with greater accuracy if $\overline{\mathrm{NAQ}}_U$ and the gender of the speaker are also included in the linear model. In contrast, the addition of $\overline{\mathrm{NAQ}}_A$ results in only a small increase in the linear fit. Stepwise linear regression, using bidirectional elimination and changes in the sum of squared errors (significance of the change was tested with F-test) as the elimination criterion, identifies the second predictor combination in Table 3 as optimal for both POT_A values and the fourth predictor combination for POT_U values. The corresponding model equations for POT_A are

$$\mathrm{POT}_A = 7.02\,\mathrm{ms} \cdot \overline{\mathrm{NAQ}}_U + 0.535 \cdot \bar{T}_o + 0.047\,\mathrm{ms} \quad (7a)$$
$$\mathrm{POT}_A = 7.02\,\mathrm{ms} \cdot \overline{\mathrm{NAQ}}_U + 0.535 \cdot \bar{T}_o + 1.33\,\mathrm{ms} \quad (7b)$$

for males and females, respectively. Similarly for POT_U,

$$\mathrm{POT}_U = 6.91\,\mathrm{ms} \cdot \overline{\mathrm{NAQ}}_U + 0.311 \cdot \bar{T}_o + 2.07\,\mathrm{ms} \quad (8a)$$

for both males and females. All of these models indicate that phonation onset occurs more slowly if the target phonation is breathier or has lower fundamental frequency.

5 Discussion

A comparative study of phonation onsets in glottal area and glottal flow has been carried out using both pulse-wise parameters and amplitude envelopes. GIF was used to estimate the glottal flow from the acoustic voice pressure signal. Despite the transient nature of phonation onsets, the produced glottal flow estimates appear reasonable when compared with the corresponding glottal area waveforms.

A wide variety of phonation onset paths was observed in the area and flow waveforms as well as in the pulse-wise parameters and amplitude envelopes. The measurement setup favored soft and breathy onsets over hard onsets for both normal and breathy target phonation types. In some samples, pulse shapes become largely constant, i.e., target

Table 3: Comparison of linear models for POT_A and POT_U with different predictors. All predictor combinations that produce a significant model for at least one of the POT values are shown.

Predictors				Model fit, POT_A			Model fit, POT_U		
$\overline{\mathrm{NAQ}}_U$	$\bar{T}_o$	gender	$\overline{\mathrm{NAQ}}_A$	R^2	F-statistic	p	R^2	F-statistic	p
x	x	x	x	0.885	9.60	0.015	0.659	2.42	0.18
x	x	x		0.883	15.1	0.0033	0.655	3.80	0.077
	x	x	x	0.777	6.97	0.022	0.578	2.74	0.14
x	x			0.734	9.64	0.0098	0.589	5.01	0.045
	x		x	0.629	5.93	0.031	0.514	3.70	0.080
	x			0.482	7.45	0.026	0.387	5.04	0.055

phonation was reached after only a couple of glottal cycles, and at the other extreme, m05 contains an extremely slow onset especially in amplitude growth. While m05 is treated as a possible outlier in this investigation, the same speaker produced two other samples (m04 and m06) which were in line with the rest of the data. Sample m05 hence simply appears to present a possible but uncommon onset control strategy.

These comparisons reveal that key features of phonation onsets appear to have a close relationship in glottal area and glottal flow. The inflection instant of the amplitude envelope, which is used to compare the timing of the onsets, occurs in glottal flow, on average, 2.3 ms after the corresponding instant in GAW. In the acoustic signal, the inflection instant occurs, on average, 2.6 ms after the corresponding instant in GAW. Paired sign tests indicate that both of these delays are statistically significant although excluding the potential outlier m05 from the test results in the delay between the inflection instants in the acoustic signal and GAW becoming nonsignificant. Since the acoustic delay of the voice signal has been compensated for, the main factors contributing to the observed delays, or lack thereof, are likely related to physiology, such as changing subglottal pressure, fluid dynamics phenomena, such as the skewing of the flow pulses, and non-linearities in the initiation of flow-induced vibrations. Unfortunately, phonation onsets were observed only in 11 samples of the 60 utterances that were recorded for the multi-channel database as described in Section 2.1. The small sample size hinders conducting powerful statistical tests, such as ANOVAs, to understand the detailed relationship between the inflection points and to explore effects of the underlying factors.

Inflection point data is not available in literature, but for comparison, Patel *et al.* (2017a) observed that first oscillations in the acoustic signal occur approximately 17 ms after the first vocal fold oscillations and approximately 6 ms before first contact of the vocal folds in men and 11 ms before the contact in women. The difference between the results using inflection points and those of Patel *et al.* (2017a) may be partially attributable to the fact that the inflection points occur later in the onset than first oscillations in both GAWs and acoustic signals, and the inflection points in the GAWs typically occur slightly before first vocal fold contact. Hence, any difference in the rate at which the amplitude envelopes grow would cause changes to the relative timings. The amplitude envelope-based measures are also more robust against noise than picking time instants manually from HSV data and acoustic signals; hence, they are less sensitive to the properties of the measurement setup and equipment (the sensitivity of the microphone, illumination of the glottis, etc.). The robustness of computing the inflection points suggests that they might be usable in onset detection. Further study is required, however, to compare this to other methods of detecting onsets, such as manual instant identification from HSV data (Patel *et al.*, 2017a), thresholding of the vibrating length of the vocal folds (Ikuma *et al.*, 2016), or automatic processing of electrolaryngography and acoustic signal pairs (D'Amario *et al.*, 2018).

The POT_A values obtained (M = 5.8 ms, range 4.6–11.0 ms excluding m05) are consistent with those reported by Patel *et al.* (2017b) ($M = 7$ ms, range 2–11 ms for men, and $M = 6$ ms, range 2–12 ms for women), even though the fitting procedure used in this work uses only the phonation initiation segment instead of the full sample. The mean POT values reported by Petermann *et al.* (2016) for the envelope fitting procedure most closely matching this study (117 ms for men and 66 ms for women) are notably higher, however. The onsets used by Petermann *et al.* (2016) appear to have long segments of amplitude growth (their Figs. 4–8), so the difference in POT values is more likely caused by differences in the speech material given to speakers ([‘mama] in Petermann *et al.* (2016), three repetitions of [hi] in Patel *et al.* (2017b), and prolonged [i] in the present study) rather than by differences in envelope fitting procedures.

There are no POT values for glottal flow or acoustic signal available in literature for comparison. However, the high correlation between POT_A and POT_U, as well as between POT_A and POT_M, are plausible, as interactions between vocal folds, glottal flow, and the vocal tract mean that changes in vocal fold oscillation amplitudes likely propagate to other parts of the speech production system as well.

POT is based on the envelope function introduced by Mergell *et al.* (1998). The mismatch between the envelope function and the HEs of glottal area and flow was observed to be largest at the end of the phonation initiation segment. The Mergell envelope assumes that amplitude growth at phonation onset follows a simple saturation pattern. This is not, however, always the case with natural speech. Instead, the fast amplitude growth of the phonation initiation segment is often followed by a segment with a slower rate of growth or a local maximum and decreasing amplitudes. Similar observations were made by Petermann *et al.* (2016) and Patel *et al.* (2017b), who fitted the Mergell envelope to segments which also contained what is, in this investigation, considered the stabilization segment, and hence observed even larger discrepancies between the Mergell envelope and the data. Despite this, the Mergell envelope remains a useful tool. Since only the phonation initiation segment was used in the fitting procedure, the envelope function covered the segment where it best describes the data. HEs of the glottal area and flow are very similar; hence, the Mergell envelopes deviate from them in a systematic manner, resulting in comparable parameters for the HEs, even if the function itself is not a perfect representation of the HE.

It was observed that the POT values depended on a combination of pulse-wise parameters of stabilized phonation and gender. The effect of increased breathiness in (7)–(8) is to increase POT, i.e., slow down the onset. This is opposite to the observation made by Kunduk *et al.* (2017), whose sole female speaker produced onsets with shorter transient durations at breathy phonation compared to normal. However, the transient duration used by Kunduk *et al.* (2017) can include transition regions before and after the phonation initiation segment used to compute POT values in this work; hence, depending on these transition regions, the relative durations of onsets may change. It is generally not surprising that the target pitch and phonation type which the speaker aims at in the stabilization segment affects how phonation is initiated. Different laryngeal posturing prior to phonation has been observed to result in different types of phonation (Shiba and Chhetri, 2016), and different pitches have been noted to be associated with, e.g., different subglottal pressures (Titze, 1989) and vocal fold lengths (Sonninen *et al.*, 1992; Riede and Brown, 2013: Fig. 4). It would be expected that the control strategy used to initiate phonation would encompass the entire phonation onset from prephonatory gestures to stable phonation, and that this controls strategy would reflect the physiological state needed to produce the target phonation.

Previous studies have found that female speakers produce, on average, smaller POT values than males (Patel *et al.*, 2017b; Petermann *et al.*, 2016). Equations (7)–(8) suggest that this is mainly due to the higher pitch of female voices, whereas at equal pitches POT values for females would be slightly higher than for males. However, the effect of gender alone (independent of pitch) observed in this study has limited generalizability as the small number of female speakers makes it impossible to separate a gender effect from the effect of a particular strategy used by the female speaker F04 who produced samples f02 and f03.

A more accurate estimation of coefficients in quantitative relationships between the different data modalities, such as (5)–(8), would require a larger number of samples. The measurement setup and procedures were not specifically designed to capture phonation onsets. However, the number of usable onset samples is comparable to the number of usable samples in Murtola *et al.* (2018), which makes use of the dataset for which the measurement setup was designed. A larger dataset would be desirable but its acquisition is time-consuming (2–3 hours per speaker) and cannot be done by increasing the number of repetitions per speaker due to the invasiveness of HSV. Results from smaller datasets, such as the those presented above, are hence vital in guiding the design of experimental setups for larger data acquisition efforts.

The two main aims of this study were to compare changes in glottal pulse shapes in glottal area and flow signals qualitatively, and to develop quantitative relationships between key parameters of amplitude envelopes of these signals. The small sample size meant that a universal description of pulse shape changes was not obtained. Yet, the large variety of parameter trajectories indicates that glottal area and flow cannot be assumed to follow completely identical onset patterns. The generalizability of the quantitative relationships obtained is also limited by the small and non-balanced dataset. The results do, however, support the baseline assumption that glottal area and flow signals carry largely identical information about the amplitude features of onsets.

6 Conclusions

A multichannel dataset, comprising synchronized high-speed videoendoscopy images and electroglottography and free-field microphone signals, was used to investigate phonation onset in vowel production in healthy adults. Qualitative comparison of the glottal area extracted from the high-speed images, and the glottal flow estimated from the microphone signal using glottal inverse filtering, revealed that the two signals are particularly similar at the beginning of the onset. Trajectories of pulse-wise parameters reveal that there is a large variety of ways in which quasi-stable phonation can be reached.

Quantitative comparisons were carried out between key parameters, point of inflection and POT, describing the amplitude envelopes of the glottal area and the corresponding parameters in the envelopes of the glottal flow and acoustic signal. Although, the quantitative results have large margins of error, they do nevertheless show that amplitude information extracted from glottal area and flow can, as a first approximation, be treated interchangeably. However, while glottal flow obtained by GIF may yield a reasonable estimate for onset parameters of the glottal area, and *vice versa*, in healthy adults, this cannot be generalized to pathological voice where GIF methods often fail.

The data also indicated that quantitative relationships between POT values and pulse-wise parameters of stabilized phonation may be achievable. Overall, the above results suggest that future research focusing on the shape of glottal area and flow pulses during and following phonation onsets may yield more information about phonation onsets as full interactive process from vocal fold vibrations to the produced voice signals.

Acknowledgments

This study was funded by the Academy of Finland (project no. 312490).

References

Airaksinen, M., Raitio, T., Story, B., and Alku, P. (**2014**). "Quasi closed phase glottal inverse filtering analysis with weighted linear prediction," IEEE/ACM Trans. Audio, Speech, Language Process. **22**(3), 596–607, 10.1109/TASLP.2013.2294585.

Airas, M. (**2008**). "TKK Aparat: An environment for voice inverse filtering and parameterization," Logoped. Phoniatr. Vocol. **33**(1), 49–64, 10.1080/14015430701855333.

Airas, M., and Alku, P. (**2006**). "Emotions in vowel segments of continuous speech: Analysis of the glottal flow using the normalised amplitude quotient," Phonetica **63**(1), 26–46, 10.1159/000091405.

Alku, P. (**1992**). "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," Speech Commun. **11**(2), 109–118, 10.1016/0167-6393(92)90005-R.

Alku, P. (**2011**). "Glottal inverse filtering analysis of human voice production - A review of estimation and parameterization methods of the glottal excitation and their applications," Sadhana **36**(5), 623–650, 10.1007/s12046-011-0041-5.

Alku, P., Bäckström, T., and Vilkman, E. (**2002**). "Normalized amplitude quotient for parametrization of the glottal flow," J. Acoust. Soc. Am. **112**(2), 701–710, 10.1121/1.1490365.

Alku, P., and Vilkman, E. (**1995**). "Effects of bandwidth on glottal airflow waveforms estimated by inverse filtering," J. Acoust. Soc. Am. **98**(2), 763–767..

Alku, P., Pohjalainen, H., and Airaksinen, M. (**2017**). "Aalto Aparat — A freely available tool for glottal inverse filtering and voice source parameterization," in *Subsidia: Tools and Resources for Speech Sciences*, Malaga, Spain, June 21-23.

Björkner, E., Sundberg, J., Cleveland, T., and Stone, E. (**2006**). "Voice source differences between registers in female musical theater singers," J. Voice **20**(2), 187–197, 10.1016/j.jvoice.2005.01.008.

Braunschweig, T., Flaschka, J., Schelhorn-Neise, P., and Döllinger, M. (**2008**). "High-speed video analysis of the phonation onset, with an application to the diagnosis of functional dysphonias," Med. Eng. Phys. **30**(1), 59–66, 10.1016/j.medengphy.2006.12.007.

Childers, D. G., and Ahn, C. (**1995**). "Modeling the glottal volume-velocity waveform for three voice types," J. Acoust. Soc. Am. **97**(1), 505–519, 10.1121/1.412276.

Childers, D. G., Naik, J. M., Larar, J. N., Krishnamurthy, A. K., and Moore, G. P. (**1985**). "Electroglottography, speech, and ultra-high speed cinematography," in *Vocal fold physiology*, edited by I. R. Titze and R. C. Scherer (The Dencer Center For The Performing Arts, Denver), pp. 202–220.

D'Amario, S., Daffern, H., and Bailes, F. (**2018**). "A new method of onset and offset detection in ensemble singing," Logop. Phoniatr. Vocology, 1–16, 10.1080/14015439.2018.1452977.

de Cheveigné, A., and Kawahara, H. (**2002**). "Yin, a fundamental frequency estimator for speech and music," J. Acoust. Soc. Am. **111**(4), 1917–1930, 10.1121/1.1458024.

Hammer, M. J. (**2013**). "Aerodynamic assessment of phonatory onset in Parkinson's disease: evidence of decreased scaling of laryngeal and respiratory control," J. Parkinsons Dis. **3**(2), 173–179.

Hertegård, S., and Gauffin, J. (**1995**). "Glottal area and vibratory patterns studied with simultaneous stroboscopy, flow glottography, and electroglottography," J. Speech Hear. Res. **38**(1), 85–100.

Holmberg, E. B., Hillman, R. E., and Perkell, J. S. (**1988**). "Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice," J. Acoust. Soc. Am. **85**(2), 511–529, 10.1121/1.396829.

Ikuma, T., Kunduk, M., Fink, D., and McWhorter, A. J. (**2016**). "A spatiotemporal approach to the objective analysis of initiation and termination of vocal-fold oscillation with high-speed videoendoscopy," J. Voice **30**(6), 756.e21 – 756.e30, 10.1016/j.jvoice.2015.09.007.

Kunduk, M., Ikuma, T., Blouin, D. C., and McWhorter, A. J. (**2017**). "Effects of volume, pitch, and phonation type on oscillation initiation and termination phases investigated with high-speed videoendoscopy," J. Voice **31**(3), 313 – 322, 10.1016/j.jvoice.2016.08.016.

Lohscheller, J., Toy, H., Rosanowski, F., Eysholdt, U., and Döllinger, M. (**2007**). "Clinically evaluated procedure for the reconstruction of vocal fold vibrations from endoscopic digital high-speed videos," Med. Image Anal. **11**(4), 400–413, 10.1016/j.media.2007.04.005.

Mergell, P., Herzel, H., Wittenberg, T., Tigges, M., and Eysholdt, U. (**1998**). "Phonation onset: Vocal fold modeling and high-speed glottography," J. Acoust. Soc. Am. **104**(1), 464–470.

Murtola, T., Alku, P., Malinen, J., and Geneid, A. (**2018**). "Parameterization of a computational physical model for glottal flow using inverse filtering and high-speed videoendoscopy," Speech Commun. **96**, 67–80, 10.1016/j.specom.2017.11.007.

Oppenheim, A. V., and Schafer, R. W. (**1989**). *Discrete-Time Signal Processing* (Prentice-Hall, Englewookd Cliffs, New Jersey).

Orlikoff, R. F., Deliyski, D. D., Baken, R., and Watson, B. C. (**2009**). "Validation of a glottographic measure of vocal attack," J. Voice **23**(2), 164–168, 10.1016/j.jvoice.2007.08.004.

Patel, R. R., Forrest, K., and Hedges, D. (**2017**a). "Relationship between acoustic voice onset and offset and selected instances of oscillatory onset and offset in young healthy men and women," J. Voice **31**(3), 389.e9–389.e17, 10.1016/j.jvoice.2016.09.021.

Patel, R. R., Walker, R., and Döllinger, M. (**2017**b). "Oscillatory onset and offset in young vocally healthy adults across various measurement methods," J. Voice **31**(4), 512.e17–512.e24, 10.1016/j.jvoice.2016.12.002.

Petermann, S., Kniesburges, S., Ziethe, A., Schützenberger, A., and Döllinger, M. (**2016**). "Evaluation of analytical modeling functions for the phonation onset process," Comput. Math. Methods Med. **2016**, 1–10. 10.1155/2016/8469139.

Riede, T., and Brown, C. (**2013**). "Body size, vocal fold length, and fundamental frequency–implications for mammal vocal communication," Nova Acta Leopold. **111**(380), 1–20.

Shiba, T. L., and Chhetri, D. K. (**2016**). "Dynamics of phonatory posturing at phonation onset," Laryngoscope **126**(8), 1837–1843, 10.1002/lary.25816.

Sonninen, A., Hurme, P., and Vilkman, E. (**1992**). "Roentgenological observations on vocal fold length-changes with special reference to register transition and open/covered voice," Scand. J. Log. Phon. **17**, 95–106, 10.3109/14015439209098719.

Sundberg, J., Fahlstedt, E., and Morell, A. (**2005**). "Effects on the glottal voice source of vocal loudness variation in untrained female and male voices," J. Acoust. Soc. Am. **117**(2), 879–885, 10.1121/1.1841612.

Švec, J. G., and Schutte, H. K. (**1996**). "Videokymography: High-speed line scanning of vocal fold vibration," J. Voice **10**(2), 201–205, 10.1016/S0892-1997(96)80047-6.

Titze, I. R. (**1989**). "On the relation between subglottal pressure and fundamental frequency in phonation," J. Acoust. Soc. Am. **85**(2), 901–906, 10.1121/1.397562.

Watson, B. C., Baken, R., and Roark, R. M. (**2016**). "Effect of voice onset type on vocal attack time," J. Voice **30**(1), 11–14, 10.1016/j.jvoice.2014.12.004.

Watson, B. C., Baken, R., Roark, R. M., Reid, S., Ribeiro, M., and Tsai, W. (**2013**). "Effect of fundamental frequency at voice onset on vocal attack time," J. Voice **27**(3), 273 – 277, 10.1016/j.jvoice.2012.12.004.

Wittenberg, T., Mergell, P., Tigges, M., and Eysholdt, U. (**1997**). "Quantitative characterization of functional voice disorders using motion analysis of high-speed video and modeling," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 3, pp. 1663–1666, 10.1109/ICASSP.1997.598831.