JASA ARTICLE

Velocity differences in laryngeal adduction and abduction gestures

Christian Kleiner,[1,a)] Marie-Anne Kainz,[2] Matthias Echternach,[2,b)] and Peter Birkholz[1]
[1]*Institute of Acoustics and Speech Communication, Technische Universität Dresden, Dresden, Germany*
[2]*Division of Phoniatrics and Pediatric Audiology, Department of Otorhinolaryngology, Munich University Hospital (LMU), Munich, Germany*

ABSTRACT:
The periodic repetitions of laryngeal adduction and abduction gestures were uttered by 16 subjects. The movement of the cuneiform tubercles was tracked over time in the laryngoscopic recordings of these utterances. The adduction velocity and abduction velocity were determined objectively by means of a piecewise linear model fitted to the cuneiform tubercle trajectories. The abduction was found to be significantly faster than the adduction. This was interpreted in terms of the biomechanics and active control by the nervous system. The biomechanical properties could be responsible for a velocity of abduction that is up to 51% higher compared to the velocity of adduction. Additionally, the adduction velocity may be actively limited to prevent an overshoot of the intended adduction degree when the vocal folds are approximated to initiate phonation. © *2022 Acoustical Society of America*.
https://doi.org/10.1121/10.0009141
(Received 16 March 2021; revised 2 December 2021; accepted 5 December 2021; published online 3 January 2022)
[Editor: Anders Lofqvist] Pages: 45–55

I. INTRODUCTION

Speech articulation is a complex process in which the nervous system actively controls the biomechanical system that interacts with the aerodynamical forces. This makes the understanding of articulator trajectories a challenging task (Fuchs and Perrier, 2005). An issue of great interest in many studies on articulation is the velocity of the articulators when they approach their targets (Kelso *et al*., 1985; Kohler, 1981; Nittrouer, 1991; Summers, 1987). Such studies with a special reference to the laryngeal articulatory function are, for example, Munhall and Ostry (1983), Cooke *et al*. (1997), Munhall *et al*. (1985), and Löfqvist and Yoshioka (1981). Some studies found that certain articulators move faster, on average, in one direction than in the other (Kelso *et al*., 1985; Kollia *et al*., 1995; Parush *et al*., 1983; Smith *et al*., 1993). These direction-dependent velocity differences could, at least partly, result from the iomechanical properties and, therefore, be intrinsic to the biomechanical system. This was suggested for the tongue (Birkholz *et al*., 2011a; Recasens and Espinosa, 2010; Thiele *et al*., 2020), the lips and jaw (Birkholz and Hoole, 2012), and the vocal fold elongation and shortening (Sundberg, 1979; Xu and Sun, 2002).

To our knowledge, the potential velocity differences between the laryngeal adduction and abduction, i.e., between the narrowing and widening of the laryngeal airway, respectively, have only been studied indirectly thus far. For example, the *vocal fold* adduction and abduction (as one discrete component in the laryngeal adduction and abduction) were studied by means of a three-dimensional biomechanical model (Hunter *et al*., 2004). This modeling study provided evidence that the geometrical structure of the biomechanical system favors "abduction over adduction in both peak speed and response time." A limitation of this model was that it did not include the differences in the contraction times of the intrinsic laryngeal muscles. Mårtensson and Skoglund (1964), Cooper *et al*. (1994), and Alipour *et al*. (2005) provided the evidence for such differences in dogs and other animals. They measured *in vitro* the time from the beginning to the peak of the contraction of four laryngeal muscles, namely, the posterior cricoarytenoid (PCA), interarytenoid (IA), thyroarytenoid (TA), and lateral cricoarytenoid (LCA) muscle. Their data suggest that the PCA and IA are slow muscles and the TA is a fast muscle, where the contraction time of a slow muscle was found to be about twice as long as that of a fast muscle. Regarding the LCA, it is hard to say whether it is a fast or slow muscle due to the contradictory observations in Mårtensson and Skoglund (1964) and Alipour *et al*. (2005), respectively. It must be mentioned that *in vivo* there might be no such clear difference between the slow and fast muscles at all as each muscle could have its own relative distribution of slow and fast muscle fibers which are activated in a task-dependent manner. Nevertheless, to assess what influence the contractile properties of the laryngeal muscles can have on the vocal fold adduction and abduction, one has to consider the primary functions of the muscles in this respect. Hirose (1976) and Hirose and Ushijima (1978) suggested that the PCA and IA functions are activated reciprocally for the vocal fold abduction and adduction, respectively. The data of Löfqvist and Yoshioka (1980) suggest that this

[a)]Electronic mail: christian.kleiner@tu-dresden.de
[b)]ORCID: 0000-0003-0095-5360.

assumption is valid for the single adduction and abduction gestures, although the picture seems to be more complicated for clusters of either of the two gestures. In a more recent study, Hillel (2001) suggested that the PCA independently abducts the vocal folds and the LCA, IA, and TA jointly adduct the vocal folds. Based on this, one can conclude that the contractile properties, i.e., the differences in the contraction times of the laryngeal muscles, lead to the vocal fold adduction being faster than the abduction, as already suspected by Stevens (1999). Note that this is contrary to what was suspected for the geometrical properties mentioned above, which raises the question of what can be expected under the consideration of both of these properties. This was performed by Titze and Hunter (2007) using a two-dimensional biomechanical model. Their data suggest that the vocal fold abduction is faster than the adduction, thereby providing evidence in the same direction as that of Hunter *et al.* (2004), where only the geometrical properties were considered. In summary, the consideration of the biomechanical properties (such as the geometrical and contractile properties) apparently suffices to make a velocity difference in the vocal fold adduction and abduction expectable. In particular, there is evidence that the vocal fold abduction is faster than the adduction and this is caused by the geometrical rather than by the contractile properties. But it is still unknown as to what extent this applies to the larger laryngeal articulatory function as the vocal fold adduction and abduction represent only one discrete component in the laryngeal adduction and abduction.

The experimental analysis of this is challenging partly because no robust approach has yet been established as a standard for the quantitative analysis of the laryngeal adduction and abduction. The present study aims to overcome this and, thus, enrich the academic discussion. To this end, the potential velocity differences between the laryngeal adduction and abduction in multiple subjects were studied by means of a new experimental paradigm using laryngoscopy. This could help to disambiguate the roles of the different biomechanical properties, such as the muscle contraction time and geometrical properties discussed above. This, in turn, could be a first step toward a (speech-related) macroscopic biomechanical characterization of the laryngeal adduction and abduction, which in the future could help to shed light on the interpretation of these speech kinematics in terms of the biomechanical properties vs active control by the nervous system vs aerodynamic forces. Last but not least, the present study could pave the way for more natural source modeling in articulatory speech synthesis.

II. MATERIALS AND METHODS

A. Data recording

The laryngeal adduction and abduction can be studied using different measurement techniques amongst which are electroglottography (Fant *et al.*, 1966; Rothenberg and Mahshie, 1988), photoglottography (Hoole, 1999; Ohala, 1966; Sonesson, 1959; Suthau *et al.*, 2016), electromyography (Faaborg-Andersen, 1957; Hirose and Gay, 1972; Weddell *et al.*, 1944), and magnetic resonance imaging (Baki *et al.*, 2017). However, the gold standard for the assessment of the laryngeal function is videoendoscopy. In the present study, as in previous investigations (Echternach *et al.*, 2017a,b, 2020), high-speed transnasal videoendoscopy (Fastcam SA-X2 480K, Photron, Tokyo, Japan) was performed using a flexible endoscope (ENF-GP, Olympus Europa SE & Co. KG, Hamburg, Germany) at a frame rate of 500 frames per second and a spatial resolution of 384×328 pixels. In this way, the laryngoscopic videos of eight male and eight female German adults (20–53 years of age) without any known speech or hearing disorders were recorded. Both genders were considered because several laryngeal articulatory parameters were shown to differ between the male and female speakers (Döllinger *et al.*, 2017; Holmberg *et al.*, 1988).

Each subject uttered 12 sequences, each of which contained 7 periodic repetitions of a contrasting segment pair. Here, a segment is understood as a laryngeal target. The sequences differed in segment pair, segment order, and speaking rate, as shown in Table I, to account for the possible effects of these factors. Regarding the segment pair, it was shown by Munhall *et al.* (1985) that there is a strong linear relationship between the maximum velocity and amplitude of the vocal fold posturing movements. The sequences used in the present study were designed in such a way that the amplitudes of the unilateral movements are constant throughout each sequence, but it can vary between sequences and speakers. Where it made sense, also, the *order* of the segments was changed to account for a possible involuntary emphasis of the start segment. This could, however, hardly be perceived by the experimenter in any of the utterances. Regarding the speaking rate, Birkholz *et al.* (2011a) and Thiele *et al.* (2020) indicate to which extent its active control can affect the intrinsic direction-dependent velocities of the articulators by the example of the tongue. In summary, there is evidence that the factors segment pair and speaking rate, as well as gender, could all influence the possible velocity difference between the laryngeal adduction and abduction.

TABLE I. The sequences that were uttered by each subject (/e̥/ means a whispered /e/).

Segment pair	Segment order	Speaking rate
/f/-/e/	/f/ - /e/ - /f/ - /e/ - ⋯	Slow
	/f/ - /e/ - /f/ - /e/ - ⋯	Fast
	/e/ - /f/ - /e/ - /f/ - ⋯	Slow
	/e/ - /f/ - /e/ - /f/ - ⋯	Fast
/v/-/f/	/v/ - /f/ - /v/ - /f/ - ⋯	Slow
	/v/ - /f/ - /v/ - /f/ - ⋯	Fast
	/f/ - /v/ - /f/ - /v/ - ⋯	Slow
	/f/ - /v/ - /f/ - /v/ - ⋯	Fast
/f/-/ʔ/	/f/ - /ʔ/ - /f/ - /ʔ/ - ⋯	Slow
	/f/ - /ʔ/ - /f/ - /ʔ/ - ⋯	Fast
/e̥/-/ʔ/	/e̥/ - /ʔ/ - /e̥/ - /ʔ/ - ⋯	Slow
	/e̥/ - /ʔ/ - /e̥/ - /ʔ/ - ⋯	Fast

https://doi.org/10.1121/10.0009141

The segment pairs were selected such that they induce pronounced laryngeal adduction and abduction and, as such, a contrast in the laryngeal articulatory function. This can be expected from the differences in the (peak) glottal areas of the segments, as given by Stevens (1999) and illustrated in Fig. 1, because these differences result from the larger laryngeal articulatory function. It must be mentioned that this is subject to uncertainties because the peak glottal areas can vary within certain ranges (gray ellipses). The slow and fast speaking rates were selected such that they can be conveniently realized by all of the speakers. They were defined as 500 and 375 ms per segment and regulated by a metronome.

The sequences were presented on a display and supportively uttered by the experimenter. The subjects were instructed to utter each sequence with flat intonation and, furthermore, to utter each segment in a sustained manner and switch between the segments at the metronome click. The two segments were not produced as syllables but as individual sounds. As an example, Fig. 2 shows the utterances /f-e-…/, /f-v-…/, /f-ʔ-…/, and /e̥-ʔ-…/ by the subject m01 at the slow speaking rate in terms of the spectrograms. In total, 192 utterances (12 sequences × 16 subjects) were recorded. Of these, 25 utterances got lost due to experiment interruption by the subject or technical failures in the data transmission. Another 13 utterances were not suitable for further analysis because of the incorrect execution or occlusion of the cuneiform tubercles by the epiglottis. Each of the remaining 154 utterances was converted from the proprietary video format to a series of TIFF images (Photron Fastcam Viewer 3, Photron Limited, Tokyo, Japan) by manually selecting the start frame and end frame in the laryngoscopic recordings.

B. Data processing

The laryngoscopic image series of each available utterance was processed in four steps. In the first step, each image was smoothed using a 5 × 5 binomial kernel to reduce the noise as well as the observed Moiré effect.

In the second step, the cuneiform tubercle trajectories were extracted automatically. Before explaining in detail how this was done, the decision for this anatomical structure will be motivated. The cuneiform tubercles are part of the aryepiglottic folds and, therefore, are directly coupled with

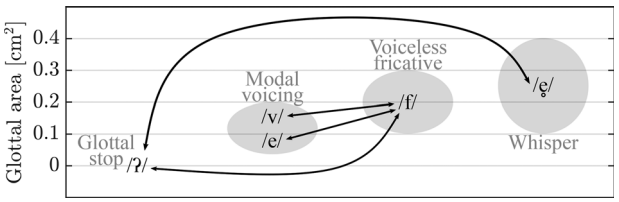

FIG. 1. The segment pairs in the continuum from the vocal fold adduction to abduction (/e̥/ means a whispered /e/). The glottal area is zero during a glottal stop (/ʔ/). The three gray ellipses schematically represent the expected ranges of the (peak) glottal area for the modal voicing (/e/ and /v/), voiceless fricatives (/f/), and whispered vowels (/e̥/), respectively (Stevens, 1999). The peak glottal area is expected to be somewhat greater during /v/ than during /e/ (Stevens, 1999).

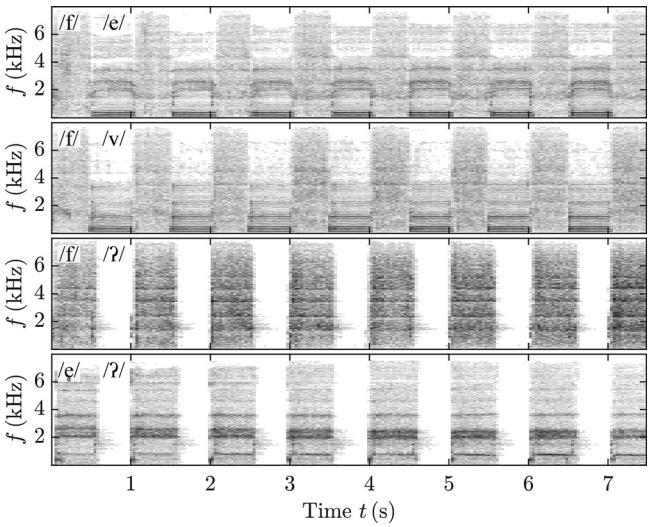

FIG. 2. The sample realizations of the sequences /f-e-…/, /f-v-…/, /f-ʔ-…/, and /e̥-ʔ-…/ (top to bottom) by the subject m01 at the slow speaking rate, displayed as spectrograms. Each spectrogram was generated with the MATLAB R2019b (The MathWorks Inc., Natick, MA) function `spectrogram( )` using Hamming windows of length 46.4 ms (2048 samples) overlapping by 11.3 ms (500 samples) with the argument `"MinThreshold"` set to −100 for visualization purposes. The sampling frequency of each audio signal was 44.1 kHz.

the arytenoid cartilages. That is, the movement of the cuneiform tubercles reflects the vocal fold adduction and abduction (Zhang, 2016) but also the movement of the aryepiglottic folds going beyond this. The latter may be the case, for example, in the articulation of /ʔ/ when the vocal folds are already fully adducted but the cuneiform tubercle approximation still continues (Esling, 1996). In contrast to other structures such as the arytenoid cartilages, the cuneiform tubercles can be directly observed in the laryngoscopic images and there are already some approaches to automatically track them in Zhuang *et al.* (2013) and Ferster *et al.* (2019). As a preparation to the tracking of the cuneiform tubercles in the present study, a bitshift by one or two positions was applied, if necessary, to each pixel of a frame to increase the brightness and contrast until both of the cuneiform tubercles were clearly visible without being saturated. The laryngoscopic images shown further below were possibly obtained using a bitshift of more positions to improve the overall visibility of all of the anatomical structures but only for visualization purposes. After this preparation, each of the two cuneiform tubercles was manually marked in terms of a rectangular region in one of the frames [see Fig. 3(a)] after the first two "training" segments. These regions were then tracked in the following frames until the end of the utterance or the tracking failed. The tracking failed only rarely but mostly due to the complete occlusion of the cuneiform tubercles by the epiglottis or a burst of saliva. Generally, the tracking of the cuneiform tubercles is a challenging task. The reasons are their low texture, weak edges, appearance change during movement and endoscopic artifacts (Ali *et al.*, 2021), and especially the specular reflections (Brelstaff and Blake, 1988; Gröger *et al.*, 2001; Ragheb and Hancock, 2003; Shah *et al.*, 2017). In the

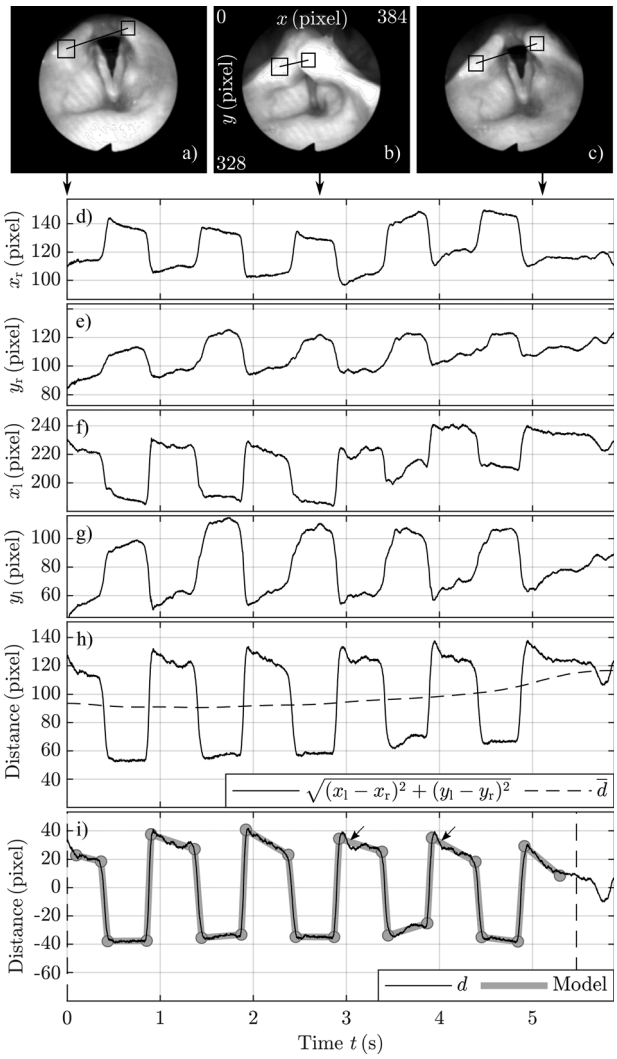

FIG. 3. The processing of the utterance /f-ʔ-…/ of the subject m01 at the slow speaking rate. The (a) manually defined regions (black rectangles) for the tracking of the left and right cuneiform tubercles, (b) tracked regions during/ʔ/, and (c) tracked regions during /f/ are displayed. The [(d),(e)] positions of the right cuneiform tubercles and [(f),(g)] positions of the left cuneiform tubercles are shown. (h) The Euclidean distance $d_{\text{tot}}(n)$ between the cuneiform tubercles with a low-frequency drift $\bar{d}(n)$ (dashed curve) and (i) "drift-free" Euclidean distance $d(n)$ without a low-frequency drift are shown. The largest contiguous section of valid windows (vertical dashed lines), obtained as explained in Fig. 5, was considered for the modeling. Corresponding to the five full periods within this section, a number of 22 points (gray markers) fully defined the model (gray curve). The optimal model is displayed in the sense of the minimum total squared error as further detailed in the text. The black arrows point to the overshoots in the laryngeal abduction.

present study, the tracking was performed automatically using an improved correlation filter (Lukežič *et al.*, 2017) based on the work of Hester and Casasent (1980) and Bolme *et al.* (2010) and implemented in the OpenCV (Bradski, 2000) 4.1 `TrackerCSRT class`. The tracking of the left and right cuneiform tubercles was checked visually and saved as the discrete-time positions $(x_l(n), y_l(n))$ and $(x_r(n), y_r(n))$, respectively, where $n = \{0, 1, \ldots, N-1\}$ is the frame index and N is the total number of frames [see Figs. 3(d)–3(g)].

In the third step, the Euclidean distance between the cuneiform tubercles d_{tot} was calculated and assumed to be the superposition of two signals, namely, the low-frequency drift $\bar{d}(n)$ of the Euclidean distance and the actual "drift-free" Euclidean distance $d(n)$ of interest,

$$d_{\text{tot}} = \sqrt{(x_l - x_r)^2 + (y_l - y_r)^2} := d + \bar{d}. \tag{1}$$

In Eq. (1), the sample index n was omitted for the sake of clarity. The Euclidean distance on the left-hand side of Eq. (1) is shown in Fig. 3(h) (solid curve) and Figs. 3(a)–3(c) (oblique lines). It is robust to the global movements like the camera motion and whole larynx motion, which means that during a joint shift of the (projected) absolute cuneiform tubercle positions, the change in the Euclidean distance was small as illustrated in Fig. 4. For example, the whole larynx motion was observed for the utterances /v-f-…/ and /f-v-…/, where the larynx took a lower position during /v/. The low-frequency drift of the Euclidean distance $\bar{d}(n)$ [dashed curve in Fig. 3(h)] was determined using a zero-phase low-pass finite impulse response (FIR) filter with a Gaussian impulse response. The filter length was empirically defined as three times the maximum period duration and, thus, was 3 s (1500 samples). The standard deviation of the Gaussian impulse response was set proportional to the filter length using the MATLAB R2019b (The MathWorks Inc., Natick, MA) function `gausswin( )` with default parameters, leading to a cutoff frequency of 0.23 Hz. The low-frequency drift $\bar{d}(n)$ was caused by a slow change in the vertical endoscope position and removed from the Euclidean distance to obtain the drift-free Euclidean distance $d(n)$ [see Fig. 3(i)], according to Eq. (1).

The fourth step was to exclude the invalid $d(n)$ sections, which could occur either due to an erroneous realization of the sequence by the subject or errors in tracking the cuneiform tubercles. The latter could occur as a result of the partial occlusion (specular reflections, epiglottis), confusion with similar anatomical structures (corniculate tubercles), or strong appearance changes. To objectively identify the valid sections, a rectangular window was shifted sample-wise across the $d(n)$ curve [see Fig. 5(b)], where the window length L is twice the period duration T dictated by the metronome ($T = 1$ s for the slow speaking rate and $T = 750$ ms for the fast speaking rate). At each window position, the normalized autocorrelation function was computed as

$$R(l) = \frac{1}{L-l}\frac{r(l)}{r(l=0)}, \tag{2}$$

where $r(l)$ is the autocorrelation function of the windowed $d(n)$ over the lag l [see the black curves in Figs. 5(d)–5(e)]. The complete positive correlation between the two periods in the window would mean that $R(l = T) = 1$. In reality, however, this ideal was never achieved not only because of the errors discussed above but also because of unavoidable variations in the timing and amplitude of the adduction and abduction movements. To allow for these variations within a certain range, the window was treated as valid when the

https://doi.org/10.1121/10.0009141

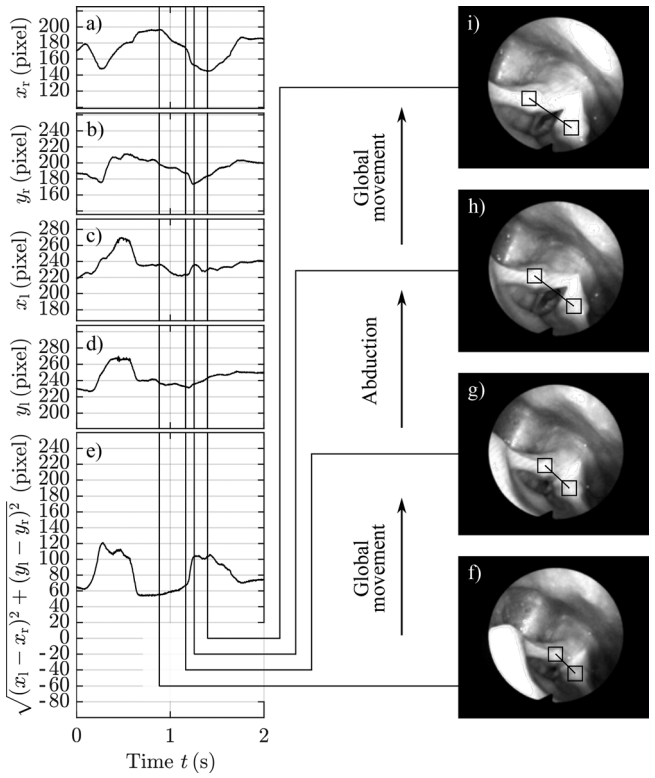

FIG. 4. The robustness of $d(n)$ as a measure for the laryngeal adduction and abduction against the global movement in the utterance /f-v-…/ of the subject m03 at the slow speaking rate. It can be seen that the global movement and laryngeal abduction can be separated from each other in $d(n)$, which is hardly possible by means of their absolute positions over time. [(a)–(d)] The absolute positions (x_r, y_r) and (x, y_l) of the right and left cuneiform tubercles, respectively, over time [see Fig. 3(b) for the definition of the absolute positions]. The (e) Euclidean distance $d_{\text{tot}}(n)$ between the cuneiform tubercles over time, [(f)–(i)] tracked regions at specific points in time, where the global movement (without the laryngeal abduction) occurs between (f) and (g) and between (h) and (i), and the laryngeal abduction occurs between (g) and (h). Here, the global movement goes along with the image scaling and translation.

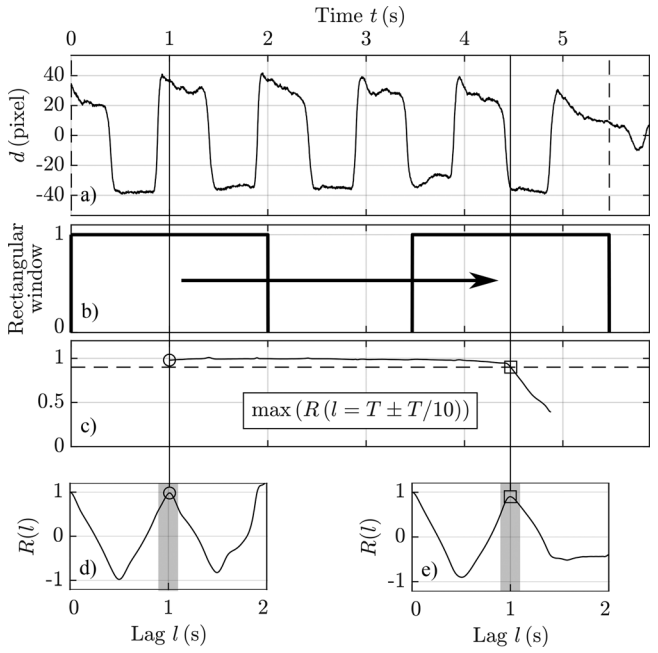

FIG. 5. The determination of the largest contiguous section of valid windows in the utterance /f-e-…/ of the subject m01 at the slow speaking rate. (a) "Drift-free" Euclidean distance $d(n)$ and largest contiguous section of valid windows (dashed lines). (b) Rectangular window shifted sample-wise across $d(n)$. The largest contiguous section of valid windows is defined by the start of the left window and the end of the right window shown here. (c) A window was considered to be valid when its $\max(R(l = T \pm T/10))$ value was greater or equal to an empirical threshold of 0.9 (dashed line). The black curve displays this value over time as the window slides over $d(n)$. The calculation of this value is illustrated in (d) and (e) for the left and right windows in (b), respectively. [(d),(e)] The normalized autocorrelation function $R(l)$ of the windowed $d(n)$ over the lag l (black curves), lag interval $l = T \pm T/10$ (gray regions), and maximum of $R(l)$ within this region, $\max(R(l = T \pm T/10))$ (markers). The markers in (d) and (e) correspond to the markers in (c).

maximum of $R(l)$ in the lag interval $l = T \pm T/10$ was greater or equal to an empirical threshold of 0.9 [see Figs. 5(c)–5(e)]. By default, the largest contiguous section of valid windows [dashed vertical lines in Fig. 5(a)] was considered for the modeling described in the following. Only in a few exceptional cases was another (smaller) contiguous section of valid windows considered.

C. Modeling

The goal was to extract the velocities of the adduction and abduction from each $d(n)$ curve, i.e., from each drift-free Euclidean distance between the cuneiform tubercles over time. To this end, a model was fitted to full periods in the considered $d(n)$ section. The general idea behind the model was that each period of $d(n)$ can be considered to consist of four phases, namely, the stationary phase of the first segment, the transition phase to the second segment, the stationary phase of the second segment, and the transition phase back to the first segment. During the stationary phase, the laryngeal configuration does not necessarily have to be static because a static (abducted) configuration is quite unusual and difficult for speakers to maintain (Löfqvist *et al.*, 1981). Instead, it can also vary as will become clearer further below. Each phase was modeled as one line piece in the (continuous) piecewise linear model, which is illustrated by the gray curve in Fig. 3(i). This model was fully defined by several points (gray markers), the number of which was obtained by multiplying the number of periods to model by the number of line segments per period (4) and adding 2, hence, 22 points were used for the 5 periods in Fig. 3(i). These points were initialized automatically and, if necessary, dragged and dropped manually. After this, the points were optimized by minimizing the total squared error between the model and the considered $d(n)$ section using the Nelder-Mead simplex method (Nelder and Mead, 1965), implemented in the MATLAB R2019b (The MathWorks Inc., Natick, MA) function `fminsearch()` until the satisfaction of the default convergence criteria. The model described above is very similar to that of Birkholz and Kleiner (2021), where it was used in another context, namely, in the investigation of the lateral pharyngeal wall movements.

The optimized model was interpreted as follows. The line pieces with even indices, i.e., the second, fourth, and so

on, represent the transition phases. Their slopes can be interpreted as the (average) velocities of the adduction and abduction, $v_{\text{adduction}}$ and $v_{\text{abduction}}$, respectively, of which one value each was extracted per modeled $d(n)$ period. For a successful extraction, however, the following three criteria had to be fulfilled. First, there had to be at least one valid window in the analyzed $d(n)$ curve. Second, in the considered $d(n)$ section, the four phases had to be clearly recognizable in at least two consecutive periods. Third, the optimized model had to fit these phases in a comprehensible way. The $d(n)$ curves of 57 utterances missed at least 1 of these criteria and were, therefore, discarded such that, together with the 38 utterances that got lost during the data recording or were discarded directly afterward (see Sec. II A), 97 of the intended 192 utterances were available for subsequent analysis.

Figure 6 shows the examples of the accepted curves and, together with Fig. 3(i), reveals some interesting details, which were explicitly not noise in the tracking of the cuneiform tubercles but actually part of their kinematics, as was verified by the visual inspection of the laryngoscopic videos. One interesting detail was a more or less pronounced overshooting in the laryngeal abduction (black arrows), which was modeled as part of the stationary phase. Here, we assumed that there *was* an underlying stationary phase, i.e., a target for the laryngeal abduction with a certain slope as is often assumed for the speech articulator movements. However, this may not necessarily be the case as a clear plateau in the abducted laryngeal state rarely occurs. The patterns indicated by the black arrows could also reflect the idea that there is no such well-defined target. Another interesting detail is the following. As already mentioned in Sec. II B, it may be the case in the articulation of /ʔ/ that the vocal folds are already fully adducted but the cuneiform tubercle approximation still continues Esling (1996). In Figs. 6(a) and 6(b), this leads to a marginal flattening or overshooting, respectively, at the end of the transition phases (white arrows). This, however, has hardly any influence on the slope of the line that represents the transition phase, which means that the laryngeal adduction velocity was largely measured as the correlated cuneiform tubercle and vocal fold movement here.

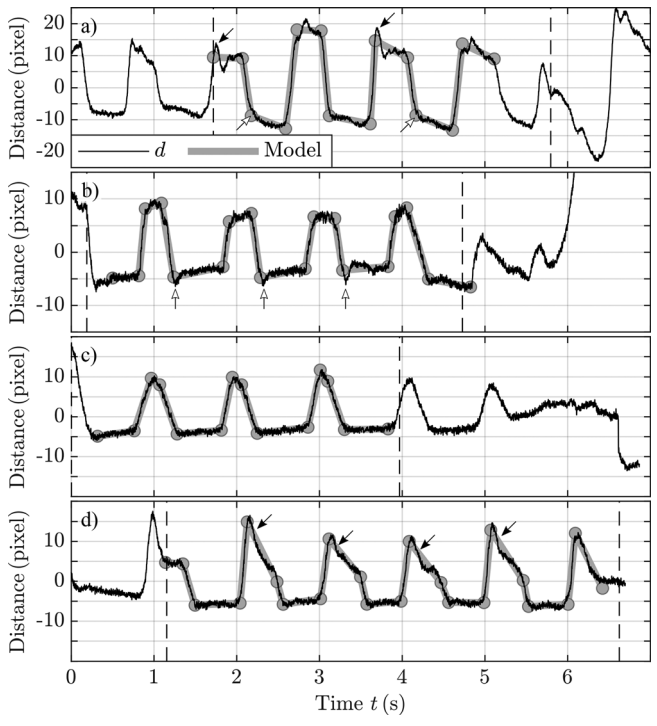

FIG. 6. The optimal models (gray curves) of the different $d(n)$ curves (black curves) in the respective largest contiguous section of valid windows (dashed lines). The (a) utterance /e̬-ʔ-…/ of the subject m05 (/e̬/ means a whispered /e/), (b) utterance /e̬-ʔ-…/ of the subject w03, (c) utterance /f-e-…/ of the subject m04, and (d) utterance /f-e-…/ of the subject w06 are depicted. All of the utterances in (a)–(d) are for the slow speaking rate. The black arrows point to the overshoots in the laryngeal abduction. The white arrows point to the marginal signal characteristics at the end of the adduction phase, namely, a flattening in (a) and an overshoot in (b).

Apart from that, one might imagine techniques other than the one presented above for the extraction of the $v_{\text{adduction}}$ and $v_{\text{abduction}}$ values. For example, one could simply pick the velocity maxima in the smoothed first derivative of the $d(n)$ curve, but this leads to the following problem. On one hand, too much smoothing reduces the edge steepness and, thus, blurs the difference between $v_{\text{adduction}}$ and $v_{\text{abduction}}$. Too little smoothing, on the other hand, does not suppress the small fluctuations, leading to more local maxima in the velocity curve and less objectivity in their selection. In addition, we often observed very linear edges (see Fig. 6), which do not lead to the pronounced maxima in the first derivative of $d(n)$ at all. These edges are, therefore, difficult to describe with the *peak* velocities often used in the speech articulation studies, but are all the better with the *average* velocities represented by the piecewise linear model. With respect to this model, the number of *four* line pieces per period was chosen deliberately. Three line pieces would be too much of an abstraction, whereas five line pieces would lead to ambiguities in their assignment to the four phases described further above. All in all, the piecewise linear model, as it was used in the present study, is considered to represent the perfect degree of abstraction and, at the same time, allows optimal preservation of the edge steepness in $d(n)$.

III. RESULTS

The results obtained from the modeling data regarding the velocity differences between the laryngeal adduction and abduction gestures are discussed further below. Before that, Fig. 7 gives an impression of the laryngeal gestures that were actually observed. It shows one representative example for each of the five segments /ʔ/, /e/, /v/, /e̬/, and /f/, where the measured peak glottal areas (white dashed contours) increase from left to right and were measured as follows. The glottal area over time for a given utterances was determined using the implementation of the seeded region growing algorithm provided by Birkholz (2016). In the resulting curve, the local maxima were selected in regions where the glottis was

https://doi.org/10.1121/10.0009141

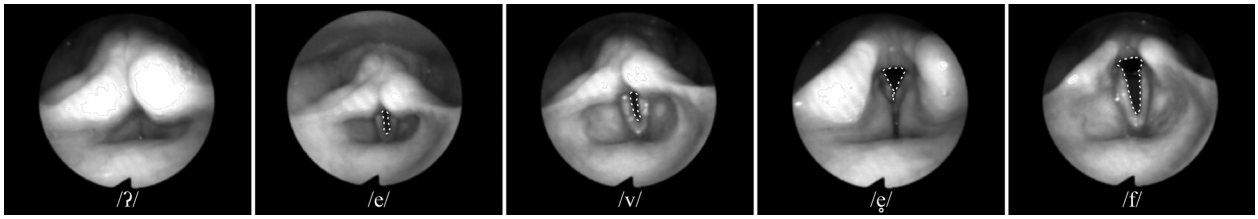

FIG. 7. The representative laryngeal configurations of all five of the segments /ʔ/, /e/, /v/, /e̥/, and /f/, where /e̥/ means a whispered /e/. The measured peak glottal areas (white dashed contours) increase from left to right. All five of the representatives were taken from the subject m01 at the slow speaking rate. More precisely, /f/ and /v/ were taken from the utterance /f-v-…/, /e/from the utterance /f-e-…/, and /e̥/ and /ʔ/ from the utterance /e̥-ʔ-…/. The visual inspection of the other utterances suggested that the laryngeal configurations shown here are representative across the segment pair, segment order, speaking rate, and subject.

not occluded by other anatomical structures such as the cuneiform tubercles. The following exemplary mean values of the peak glottal area were measured for the subject m01 at the slow speaking rate: 1138 pixels for /f/ and 314 pixels for /e/ in the utterance /f-e-…/, 1765 pixels for /f/ and 376 pixels for /v/ in the utterance /f-v-…/, and 1175 pixels for /e̥/ in the utterance /e̥-ʔ-…/. Note that the value measured for /f/ depends on the utterance, probably because of the different distances between the laryngoscope and glottis during /f-e-…/ and /f-v-…/. This is suggested by Fig. 7, where the anatomical structures and, as such, also the glottis appear smaller during /e/ than during /v/. Sudden changes in the unknown distance between the laryngoscope and glottis, as well as the limited spatial and temporal resolution of the laryngeal videos, made it difficult to determine the peak glottal area in an exact and comparable manner. Despite all of that, the above analysis shows that the initial considerations in Sec. II A (see Fig. 1) were a suitable, albeit rough, indication of which segment pairs result in the measurable mediolateral cuneiform tubercle movements.

The data obtained from the modeling consisted of 2–11 pairs of $v_{\text{adduction}}$ and $v_{\text{abduction}}$ values per utterance, where the most frequent case is two value pairs, the average is about three value pairs, and the total number is 310 value pairs. Figure 8 shows the mean $v_{\text{adduction}}$ (black-framed bars) and $v_{\text{abduction}}$ (gray bars) values for each utterance, speaking rate, and subject. It can be seen that $v_{\text{abduction}}$ is greater than $v_{\text{adduction}}$, i.e., the laryngeal abduction is faster than the adduction in most cases. Furthermore, it can be seen that the subject m01 shows overall greater velocity values than the other subjects. The reason for this is that for the subject m01, the laryngoscope was in a general lower position relative to the cuneiform tubercles than for the other subjects. The lower laryngoscope position led to a greater apparent cuneiform tubercle movement. Consequently, in the following two statistical analyses across all of the subjects, the absolute $v_{\text{adduction}}$ and $v_{\text{abduction}}$ values were not analyzed, but their ratio was analyzed.

The potential factors to analyze were the gender, segment pair, and speaking rate (see Sec. II A). Here, only the effect of the speaking rate was analyzed for two reasons. First, due to the stringent quality criteria for the recorded sequences, there was simply not enough data available to

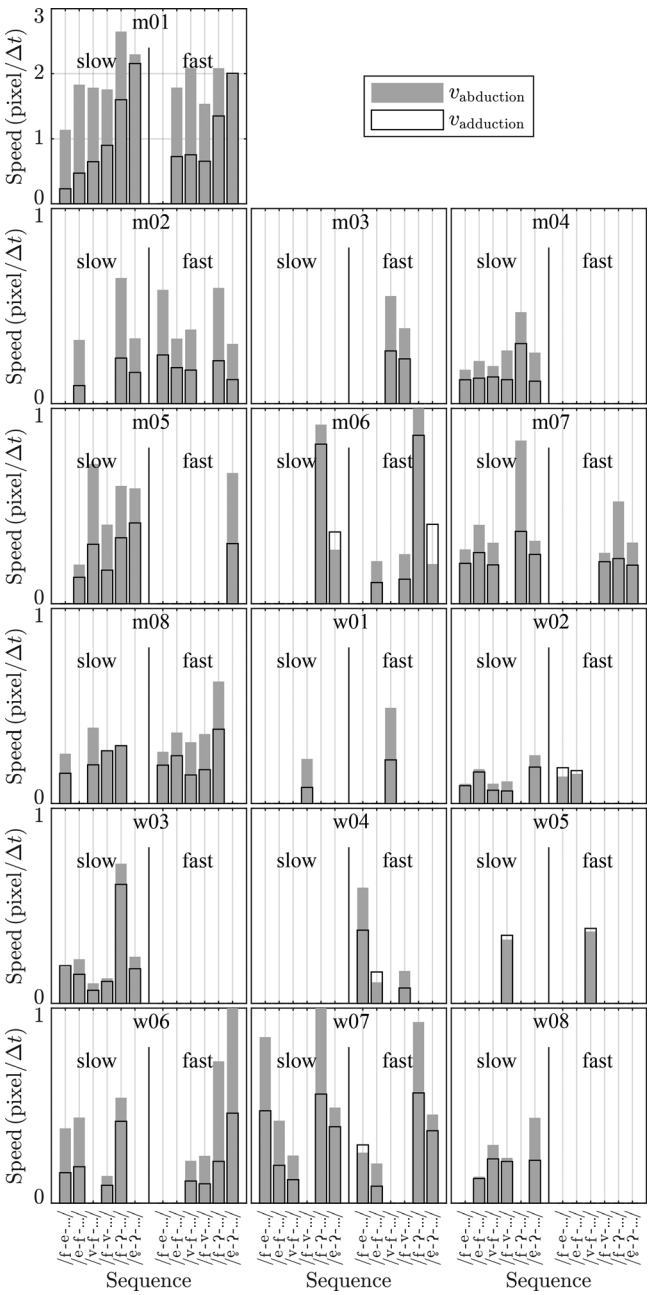

FIG. 8. The overview of the mean $v_{\text{adduction}}$ (black-framed bars) and $v_{abduction}$ (gray bars) values for all of the subjects (boxes), all of the sequences, and both speaking rates (/e̥/ means a whispered /e/).

allow more sophisticated statistical procedures to be used. Second, the factor speaking rate was considered to be particularly important because its nonsignificant effect would provide strong evidence for the hypothesis that a possible velocity difference in the laryngeal adduction and abduction is caused by biomechanical properties rather than active control.

In the first analysis, a two-sample t-test with the speaking rate as the independent variable and the decadic logarithm of $v_{\text{abduction}}/v_{\text{adduction}}$ as the dependent variable was performed. The logarithm was required to transform the distribution of the dependent variable such that it better fits the normal distribution assumed by the t-test. The normal distribution after the logarithmic transformation was checked with the help of the quantile-quantile plots. The analysis showed that the speaking rate had no significant effect. This means that for the slow speaking rate, the difference between $v_{\text{abduction}}$ and $v_{\text{adduction}}$ was similarly pronounced as for the fast speaking rate. In the second analysis, a right-tailed one-sample t-test with the same dependent variable as in the first analysis but pooled across both speaking rates was performed. The analysis showed that the mean (median) $v_{\text{abduction}}/v_{\text{adduction}}$ value of 1.89 (1.70) was significantly larger than unity ($p < 0.001$; see Fig. 9). Hence, based on the median value, the laryngeal adduction gestures took 70% longer than the abduction gestures when analyzed across all subjects, all utterances, and both speaking rates.

To corroborate the results, the manual marking of the cuneiform tubercles was performed a second time for a subset of the available utterances, namely, all six utterances of the subject m04 at the slow speaking rate. For the Euclidean distance between the cuneiform tubercles $d_{\text{tot}}(n)$, Pearson's correlation coefficient between both markings was in the range of 0.979–0.999 across all six sequences. This means that both markings led to almost identical $d_{\text{tot}}(n)$ curves. Accordingly, the mean (median) $v_{\text{abduction}}/v_{\text{adduction}}$ values differed by only 1.2% (2.9%) between both markings and, furthermore, the difference was found to be not significant. It can, therefore, be assumed that the repeated manual marking leads to similar overall results. This also proves that the piecewise linear model is robust against small fluctuations of $d(n)$.

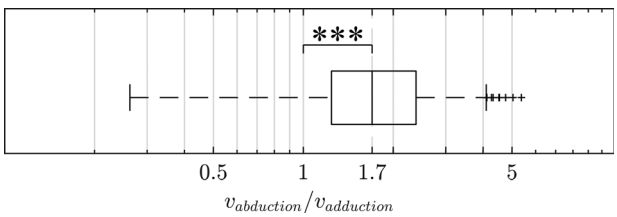

FIG. 9. The distribution of $v_{\text{adduction}}/v_{\text{abduction}}$ across all of the subjects, all of the utterances, and both speaking rates together with the results of a right-tailed one-sample t-test with the null hypothesis such that the data come from a normal distribution with the mean equal to one and an unknown variance. The MATLAB R2019b (The MathWorks Inc., Natick, MA) function `boxplot( )` was used with default parameters. $^{***}p < 0.001$.

IV. DISCUSSION AND CONCLUSIONS

For the analyzed utterances, the laryngeal abduction, i.e., the widening of the laryngeal airway, was found to be significantly faster than the adduction, i.e., the narrowing of the laryngeal airway. This direction-dependent velocity difference may be intrinsic to the biomechanical system as suggested by several studies for other articulators (Birkholz and Hoole, 2012; Birkholz *et al.*, 2011a; Recasens and Espinosa, 2010; Sundberg, 1979; Thiele *et al.*, 2020; Xu and Sun, 2002). For the laryngeal adduction and abduction, in particular, there is some scattered evidence from the modeling studies (Hunter *et al.*, 2004; Titze and Hunter, 2007) and *in vitro* studies (Alipour *et al.*, 2005; Cooper *et al.*, 1994; Mårtensson and Skoglund, 1964). These studies even provide hints about which biomechanical properties, in particular, may lead to the observed velocity difference, namely, that the effect of the geometrical properties dominates the somewhat contrary effect of the contractile properties (see Sec. I). Also, the fact that the speaking rate had no significant effect on the velocity difference in the present study agrees well with the possible biomechanical causes, although only under the assumption that the biomechanical properties, especially the possibly dominating geometrical properties, are independent of the speaking rate to some degree.

The question arises as to what could account for the possible existence of such a biomechanical mechanism that favors the fast opening of the laryngeal airway, i.e., in a way in which this could be beneficial for humans. One answer is provided by the theory of the human larynx evolution (Fink, 1974a,b, 1975). According to this, the driving force in the evolution of the larynx and its complex vagus nerve was not the folding of the laryngeal structures but rather their unfolding as for a spring recoiling to open the airway. In this view, the rapid laryngeal airway opening facilitated both the sprinting and long-distance endurance running, whereas the speech *per se* was not a determining factor and neither was holding the airway shut in a manner to allow for more efficient bracing of the arms for the upper-body strength (Esling *et al.*, 2019). The way that the laryngeal articulator works makes it logically more plausible that the observed velocity differences are caused by the geometrical properties rather than the contractile properties. The individual muscle contractions would likely not be efficient or coordinated enough in serving the mechanism's needs, especially for rapid unfolding.

Apart from the biomechanical explanation, there is at least one more possible explanation, namely, that the articulator trajectories, in general, are actively controlled by the nervous system as was suggested by Löfqvist and Gracco (2002) for the tongue. This explanation is also suitable for the observed velocity difference in the present study as will be explained herein. One could assume that the adduction for /e/ and /v/ has to be controlled more carefully than the abduction for /f/ and /e̞/ for the following reasons. On one hand, the adduction for /e/ has to be carefully controlled to

https://doi.org/10.1121/10.0009141

achieve self-sustained oscillation of the vocal folds but avoid glottalization or a glottal stop due to a target overshoot. In this context, a target overshoot means that the laryngeal adduction degree approaches the intended degree but initially exceeds it and then gradually adjusts to it, which is similar to the step response of an underdamped system. On the other hand, the adduction for /v/ has to be carefully controlled because the voiced fricatives are a difficult-to-produce class of consonants (Elie and Laprie, 2017; Ohala, 1983). To enable this careful control and thereby support the speech intelligibility, the adduction velocity during /e/ and /v/ could be actively limited, whereas careful control may be less critical for the abduction. The idea that the adduction for /e/ and /v/ must be controlled more carefully than the abduction is corroborated by the overshoots, which were widely observed for the abduction [see the black arrows in Figs. 3(i), 6(a), and 6(d)] but not for /e/ and /v/. However, this explanation is limited to the sequences containing /e/ and /v/ and cannot be applied straightforwardly to the other sequences containing /ʔ/ because a glottal stop does not require the same precision as /e/ and /v/.

For a more detailed analysis of this, the available utterances were divided into two subsets with /f-e-…/, /e-f-…/, /v-f-…/, and /f-v-…/ as the first subset, and /f-ʔ-…/ and /e̥-ʔ-···/ as the second. Each subset was analyzed in the same manner as described in Sec. III. Again, the speaking rate was found to have no significant effect for either subset. Hence, the utterances of each subset were pooled across all of the subjects and both speaking rates. The mean (median) $v_{\text{abduction}}/v_{\text{adduction}}$ value was found to be 1.99 (1.79) for the first subset, 1.71 (1.51) for the second subset, and significantly larger than unity ($p < 0.001$) for both of them.

This led us to the final assumption, which integrates both the biomechanical properties and active control by the nervous system into the explanation of the observed velocity difference between the laryngeal adduction and abduction. The biomechanical properties could be responsible for an abduction up to 51% faster compared to the adduction. On top of this, the adduction velocity may be actively limited, if necessary, leading to a 70% faster abduction compared to the adduction. This results from the median values given above and, to reemphasize what was already discussed further above, from the assumption that an active velocity limitation is *not* required for the abduction gestures, in general, and the adduction during /ʔ/ but that it *is* required for the adduction during /e/ and /v/.

A. Limitations and future directions

The present study was limited in several ways, which are discussed next. First, various quantities are conceivable to measure the laryngeal adduction and abduction, amongst which are the cuneiform tubercle positions as used in the present study, and also others such as the glottal area. The analysis of the glottal area, although conveniently possible using the software provided by Birkholz (2016), was hardly possible in many laryngoscopic videos in which the anterior part of the glottis was occluded by the epiglottis. In contrast to this, the cuneiform tubercles were visible throughout almost all of the utterances.

Second, only a limited number of contrasting segment pairs was used to induce the pronounced laryngeal adduction and abduction. In particular, the laryngeal abduction targets were defined using the segments /f/ and /e̥/ (whispered /e/) but no plosives. This might bias the present findings toward faster laryngeal abduction than adduction as the data by Löfqvist and Yoshioka (1981) suggest. They found that the maximum vocal fold abduction velocity is higher for the fricatives than for the plosives, which can be explained by different aerodynamic requirements.

Third, the sequences were only spoken egressively, and not ingressively, as the latter is difficult for many subjects. Comparing the cuneiform tubercle movements between the two ways of speaking would be interesting because the differences in the movements should be observable if they were affected by the aerodynamic factors, as demonstrated by Hoole *et al.* (1998) in another context, namely, in the investigation of the tongue body trajectories. Some evidence about the minor influence of the aerodynamic factors in the present study is provided by Löfqvist and Yoshioka (1980) and Löfqvist and Yoshioka (1981). They observed that the obstruents in a vowel context were accompanied by distinct PCA and IA activity patterns. This suggests that the observed laryngeal movements were caused by the muscular rather than by the aerodynamical forces, at least in the utterances /f-e-…/ and /e-f-…/ from the present study.

Fourth, only 97 of the intended 192 utterances (51%) were available for the final analysis. Apart from the technical failures in the data transmission, mainly, this had the following reasons. The recording of the laryngoscopic videos may involve considerable discomfort for the subject and can lead to experiment cancellation. Furthermore, the automatic tracking of the cuneiform tubercles in these videos is a challenging task for which no robust approach has yet been established as a standard. With the novel approach used in the present study, errors in the tracking of the cuneiform tubercles are, therefore, to be expected. Moreover, the sequences, which were specifically designed to unveil the intrinsic velocity differences between the laryngeal adduction and abduction, may appear artificial or unnatural to the speakers because they can hardly be interpreted as sound sequences of German and both dictated speaking rates may deviate from the speaker's individual speaking rate. This may lead to errors in the realization of the sequence, which, together with the possible errors in tracking mentioned above, can lead to the partial or complete discarding of the utterance.

Fifth, although the overall laryngeal abduction was found to be faster than the adduction, their absolute and relative values varied considerably with the subject, sequence, and speaking rate (see Fig. 8). One explanation for this is that the laryngeal targets can differ according to the language, dialect, social group, and individual and may be further influenced by the artificial speech task discussed

above. Apart from this, the observed variance may be partly explained by the factors that could not be analyzed in the present study. Regardless of the possible explanations for the observed variance, it, in fact, limits the statistical interpretability of the data to some extent. As an example, the discussion in Sec. IV could lead to the assumption that a faster speaking rate leads to a greater velocity increase in the laryngeal abduction than in the (possibly actively limited) adduction, and the speaking rate should, therefore, have a significant effect on the velocity difference between the laryngeal adduction and abduction. This effect may not be detectable simply due to the relatively large variance. But there are also other possibilities, e.g., that a faster speaking rate is not achieved by a faster adduction and abduction but by shortening of the stationary phases. However, it was not possible to reliably assess this on the basis of the available utterances. This was, in any case, beyond the scope of the present study.

Looking forward, there are at least four possible future directions. First, a follow-up study would be to investigate the velocity differences between the laryngeal adduction and abduction in more natural speech. Such a study would also aim to include more laryngeal targets, e.g., by also including plosives or recruiting speakers of other languages or dialects. Second, the tracking of the cuneiform tubercles in the laryngoscopic images developed here might be useful to study the differences in the articulatory posture between the contrasting phonation types as the measurements of this kind have not yet been applied with that scope. Beyond this, the measurement algorithms have the potential to track and quantify the laryngeal structure movements (also possibly muscle actions) for various purposes in descriptive phonetics and otorhinolaryngology. The first and second future directions may require or benefit from extensions of the tracking algorithm, which allow the movement of the cuneiform tubercles to be tracked in more planes and, even in the case of occlusion, by other anatomical structures. Third, the results could be incorporated into models for articulatory speech synthesis (Birkholz, 2013). Taking intrinsic direction-dependent velocity differences into account may lead to a more realistic movement of articulators when they approach their targets (Birkholz *et al.*, 2011b). Fourth, the results could help to disentangle the interpretation of kinematic speech data in terms of the biomechanical properties vs active control by the nervous system vs aerodynamical forces.

ACKNOWLEDGMENTS

This study was supported by the Deutsche Forschungsgemeinschaft (Grant No. BI 1639/4-1).

Ali, S., Dmitrieva, M., Ghatwary, N., Bano, S., Polat, G., Temizel, A., Krenzer, A., Hekalo, A., Guo, Y. B., Matuszewski, B., Gridach, M., Voiculescu, I., Yoganand, V., Chavan, A., Raj, A., Nguyen, N. T., Tran, D. Q., Huynh, L. D., Boutry, N., Rezvy, S., Chen, H., Choi, Y. H., Subramanian, A., Balasubramanian, V., Gao, X. W., Hu, H., Liao, Y., Stoyanov, D., Daul, C., Realdon, S., Cannizzaro, R., Lamarque, D., Tran-Nguyen, T., Bailey, A., Braden, B., East, J. E., and Rittscher, J. (**2021**). "Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy," Med. Image Anal. **70**, 102002.

Alipour, F., Titze, I. R., Hunter, E., and Tayama, N. (**2005**). "Active and passive properties of canine abduction/adduction laryngeal muscles," J. Voice **19**(3), 350–359.

Baki, M. M., Menys, A., Atkinson, D., Bassett, P., Morley, S., Beale, T., Sandhu, G., Naduvilethil, G., Stevenson, N., Birchall, M. A., and Punwani, S. (**2017**). "Feasibility of vocal fold abduction and adduction assessment using cine-MRI," Eur. Radiol. **27**(2), 598–606.

Birkholz, P. (**2013**). "Modeling consonant-vowel coarticulation for articulatory speech synthesis," PloS One **8**(4), e60603.

Birkholz, P. (**2016**). "GlottalImageExplorer—An open source tool for glottis segmentation in endoscopic high-speed videos of the vocal folds," Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung (ESSV), pp. 39–44.

Birkholz, P., and Hoole, P. (**2012**). "Intrinsic velocity differences of lip and jaw movements: Preliminary results," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 2017–2020.

Birkholz, P., Hoole, P., Kröger, B. J., and Neuschaefer-Rube, C. (**2011a**). "Tongue body loops in vowel sequences," in *Proceedings of the International Seminar on Speech Production (ISSP)*, pp. 203–210.

Birkholz, P., and Kleiner, C. (**2021**). "Velocity differences between velum raising and lowering movements," in *Proceedings of the International Conference on Speech and Computer (SPECOM)*, pp. 70–80.

Birkholz, P., Kroger, B. J., and Neuschaefer-Rube, C. (**2011b**). "Model-based reproduction of articulatory trajectories for consonant-vowel sequences," IEEE Trans. Audio Speech Lang. Proc. **19**(5), 1422–1433.

Bolme, D. S., Beveridge, J. R., Draper, B. A., and Lui, Y. M. (**2010**). "Visual object tracking using adaptive correlation filters," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2544–2550.

Brelstaff, G., and Blake, A. (**1988**). "Detecting specular reflections using lambertian constraints," in *Proceedings of the International Conference on Computer Vision (ICCV)*, pp. 297–302.

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

Cooper, D. S., Shindo, M., Hast, M. H., Sinha, U., and Rice, D. H. (**1994**). "Dynamic properties of the posterior cricoarytenoid muscle," Ann. Otolaryngol. **103**(12), 937–944.

Döllinger, M., Gómez, P., Patel, R. R., Alexiou, C., Bohr, C., and Schützenberger, A. (**2017**). "Biomechanical simulation of vocal fold dynamics in adults based on laryngeal high-speed videoendoscopy," PloS One **12**(11), e0187486.

Echternach, M., Burk, F., Burdumy, M., Herbst, C. T., Köberlein, M. C., Döllinger, M., and Richter, B. (**2017a**). "The influence of vocal mass lesions on the passaggio region of professional singers," Laryngoscope **127**(6), 1392–1401.

Echternach, M., Burk, F., Köberlein, M., Herbst, C. T., Döllinger, M., Burdumy, M., and Richter, B. (**2017b**). "Oscillatory characteristics of the vocal folds across the tenor passaggio," J. Voice **31**(3), 381.e5–381.e14.

Echternach, M., Raschka, J., Kuranova, L., Köberlein, M., Richter, B., Döllinger, M., and Kainz, M.-A. (**2020**). "Immediate effects of water resistance therapy on patients with vocal fold mass lesions," Eur. Arch. Otorhinolaryngol. **277**(7), 1995–2003.

Elie, B., and Laprie, Y. (**2017**). "Acoustic impact of the gradual glottal abduction degree on the production of fricatives: A numerical study," J. Acoust. Soc. Am. **142**(3), 1303–1317.

Esling, J. H. (**1996**). "Pharyngeal consonants and the aryepiglottic sphincter," J. Int. Phon. Assoc. **26**(2), 65–88.

Esling, J. H., Moisik, S. R., Benner, A., and Crevier-Buchman, L. (**2019**). "Laryngeal articulation and voice quality in sound change, language ontogeny and phylogeny," in *Cambridge Studies Linguistics*, edited by P. Austin, J. Bresnan, B. Comrie, S. Crain, W. Dressler, C. J. Ewen, R. Lass, D. Lightfoot, K. Rice, I. Roberts, S. Romaine, and N. V. Smith, (Cambridge University Press, Cambridge, UK), pp. 239–256.

Faaborg-Andersen, K. (**1957**). "Electromyographic investigation of intrinsic laryngeal muscles in humans," Acta Physiol. Scand. **41**(140), 1–150.

Fant, G., Ondráckova, J., Lindqvist-Gauffin, J., and Sonesson, B. (**1966**). "Electrical glottography," STL-QPSR **7**(4), 015–021, available at

https://www.speech.kth.se/prod/publications/files/qpsr/1966/1966_7_4_015-021.pdf (Last viewed 21 December 2021).
Ferster, A. P., Ferster, M. C., II, Glatthorn, H., Bacak, B. J., and Sataloff, R. T. (**2019**). "Detection of arytenoid dislocation using pixel-valued cuneiform movement," J. Voice **33**(3), 370–374.
Fink, B. R. (**1974a**). "Folding mechanism of the human larynx," Acta Oto-Laryngol. **78**(1-6), 124–128.
Fink, B. R. (**1974b**). "Spring mechanisms in the human larynx," Acta Oto-Laryngol. **77**(1-6), 295–304.
Fink, B. R. (**1975**). *The Human Larynx: A Functional Study* (Raven Press, New York).
Fuchs, S., and Perrier, P. (**2005**). "On the complex nature of speech kinematics," ZASPiL **42**, 137–165.
Gröger, M., Sepp, W., Ortmaier, T., and Hirzinger, G. (**2001**). "Reconstruction of image structure in presence of specular reflections," in *Proceedings of the DAGM German Conference on Pattern Recognition (GCPR)*, pp. 53–60.
Hester, C. F., and Casasent, D. (**1980**). "Multivariant technique for multiclass pattern recognition," Appl. Opt. **19**(11), 1758–1761.
Hillel, A. D. (**2001**). "The study of laryngeal muscle activity in normal human subjects and in patients with laryngeal dystonia using multiple fine-wire electromyography," Laryngoscope **111**(S97), 1–47.
Hirose, H. (**1976**). "Posterior cricoarytenoid as a speech muscle," Ann. Otolaryngol. **85**(3), 334–342.
Hirose, H., and Gay, T. (**1972**). "The activity of the intrinsic laryngeal muscles in voicing control," Phonetica **25**(3), 140–164.
Hirose, H., and Ushijima, T. (**1978**). "Laryngeal control for voicing distinction in Japanese consonant production," Phonetica **35**(1), 1–10.
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. **84**(2), 511–529.
Hoole, P., Gobl, C., and Chasaide, A. N. (**1999**). "Techniques for Investigating Laryngeal Articulation. Section A: Investigation of the Devoicing Gesture," edited by W. J. Hardcastle and N. Hewlett, in *Cambridge Studies Speech Science Communication* (Cambridge University Press, Cambridge, UK), pp. 294–321.
Hoole, P., Munhall, K., and Mooshammer, C. (**1998**). "Do airstream mechanisms influence tongue movement paths?," Phonetica **55**(3), 131–146.
Hunter, E. J., Titze, I. R., and Alipour, F. (**2004**). "A three-dimensional model of vocal fold abduction/adduction," J. Acoust. Soc. Am. **115**(4), 1747–1759.
Kelso, J. S., Vatikiotis-Bateson, E., Saltzman, E. L., and Kay, B. (**1985**). "A qualitative dynamic analysis of reiterant speech production: Phase portraits, kinematics, and dynamic modeling," J. Acoust. Soc. Am. **77**(1), 266–280.
Kohler, K. J. (**1981**). "Timing of articulatory control in the production of plosives," Phonetica **38**(1-3), 116–125.
Kollia, H. B., Gracco, V. L., and Harris, K. S. (**1995**). "Articulatory organization of mandibular, labial, and velar movements during speech," J. Acoust. Soc. Am. **98**(3), 1313–1324.
Löfqvist, A., Baer, T., and Yoshioka, H. (**1981**). "Scaling of glottal opening," Phonetica **38**(5-6), 265–276.
Löfqvist, A., and Gracco, V. L. (**2002**). "Control of oral closure in lingual stop consonant production," J. Acoust. Soc. Am. **111**(6), 2811–2827.
Löfqvist, A., and Yoshioka, H. (**1980**). "Laryngeal activity in swedish obstruent clusters," J. Acoust. Soc. Am. **68**(3), 792–801.
Löfqvist, A., and Yoshioka, H. (**1981**). "Interarticulator programming in obstruent production," Phonetica **38**(1-3), 21–34.
Lukežič, A., Vojíř, T., Čehovin Zajc, L., Matas, J., and Kristan, M. (**2017**). "Discriminative correlation filter with channel and spatial reliability," in *Proc. IEEE Computer Soc. Conf. Computer Vision Pattern Recog.*, pp. 6309–6318.
Mårtensson, A., and Skoglund, C. (**1964**). "Contraction properties of intrinsic laryngeal muscles," Acta Physiol. Scand. **60**(4), 318–336.
Munhall, K. G., and Ostry, D. J. (**1983**). "Ultrasonic measurement of laryngeal kinematics," in *Proceedings of the Vocal Fold Physiology Conference*, edited by I. R. Titze and R. C. Scherer, pp. 145–162.
Munhall, K. G., Ostry, D. J., and Parush, A. (**1985**). "Characteristics of velocity profiles of speech movements," J. Exp. Psychol. Human **11**(4), 457–474.
Nelder, J. A., and Mead, R. (**1965**). "A simplex method for function minimization," Comput. J. **7**(4), 308–313.
Nittrouer, S. (**1991**). "Phase relations of jaw and tongue tip movements in the production of VCV utterances," J. Acoust. Soc. Am. **90**(4), 1806–1815.
Ohala, J. (**1966**). "A new photoelectric glottograph," UCLA Working Papers in Phonetics (WPP), https://escholarship.org/uc/uclaling_wpp (Last viewed: 21.12.2021), Vol. 4, pp. 40–52.
Ohala, J. J. (**1983**). "The origin of sound patterns in vocal tract constraints," in *The Production of Speech* (Springer, New York, 1983), pp. 189–216.
Parush, A., Ostry, D. J., and Munhall, K. G. (**1983**). "A kinematic study of lingual coarticulation in vcv sequences," J. Acoust. Soc. Am. **74**(4), 1115–1125.
Ragheb, H., and Hancock, E. R. (**2003**). "A probabilistic framework for specular shape-from-shading," Pattern Recognit. **36**(2), 407–427.
Recasens, D., and Espinosa, A. (**2010**). "Lingual kinematics and coarticulation for alveolopalatal and velar consonants in Catalan," J. Acoust. Soc. Am. **127**(5), 3154–3165.
Rothenberg, M., and Mahshie, J. J. (**1988**). "Monitoring vocal fold abduction through vocal fold contact area," J. Speech Hear. Res. **31**(3), 338–351.
Shah, S. M. A., Marshall, S., and Murray, P. (**2017**). "Removal of specular reflections from image sequences using feature correspondences," Mach. Vision Appl. **28**(3-4), 409–420.
Smith, C. L., Browman, C. P., McGowan, R. S., and Kay, B. (**1993**). "Extracting dynamic parameters from speech movement data," J. Acoust. Soc. Am. **93**(3), 1580–1588.
Sonesson, B. (**1959**). "A method for studying the vibratory movements of the vocal cords," J. Laryngol. Otol. **73**(11), 732–737.
Stevens, K. N. (**1999**). *Acoustic Phonetics* (MIT Press, Cambridge, MA).
Summers, W. V. (**1987**). "Effects of stress and final-consonant voicing on vowel production: Articulatory and acoustic analyses," J. Acoust. Soc. Am. **82**(3), 847–863.
Sundberg, J. (**1979**). "Maximum speed of pitch changes in singers and untrained subjects," J. Phonetics **7**(2), 71–79.
Suthau, E., Birkholz, P., Mainka, A., and Simpson, A. P. (**2016**). "Non-invasive photoglottography for use in the lab and the field," in *Proceedings of the ITG Conference on Speech Communication*, pp. 273–277.
Thiele, C., Mooshammer, C., Belz, M., Rasskazova, O., and Birkholz, P. (**2020**). "An experimental study of tongue body loops in v1-v2-v1 sequences," J. Phonetics **80**, 100965.
Titze, I. R., and Hunter, E. J. (**2007**). "A two-dimensional biomechanical model of vocal fold posturing," J. Acoust. Soc. Am. **121**(4), 2254–2260.
Weddell, G., Feinstein, B., and Pattle, R. (**1944**). "The electrical activity of voluntary muscle in man under normal and pathological conditions," Br. J. Neurol. **67**(3), 178–257.
Xu, Y., and Sun, X. (**2002**). "Maximum speed of pitch change and how it may relate to speech," J. Acoust. Soc. Am. **111**(3), 1399–1413.
Zhang, Z. (**2016**). "Mechanics of human voice production and control," J. Acoust. Soc. Am. **140**(4), 2614–2635.
Zhuang, P., Nemcek, S., Surender, K., Hoffman, M. R., Zhang, F., Chapin, W. J., and Jiang, J. J. (**2013**). "Differentiating arytenoid dislocation and recurrent laryngeal nerve paralysis by arytenoid movement in laryngoscopic video," Otolaryngol. Head Neck Surg. **149**(3), 451–456.