COMPARISON OF OBJECT TRACKING ALGORITHMS FOR LARYNX PHANTOM MOVEMENTS IN ULTRASOUND VIDEOS

Christian Kleiner, Peter Birkholz

Institute of Acoustics and Speech Communication, Technische Universität Dresden, Dresden, Germany
christian.kleiner@tu-dresden.de

Abstract: A larynx phantom based on gelatine and an excised pig larynx was built. Larynx phantom movement was determined in RGB video for reference. Additionally, larynx phantom movement was estimated in ultrasound video using two different tracking algorithms based on optical flow and template matching, respectively. In ultrasound video, high tracking accuracies of up to 97 % were found for both tracking algorithms. This supports the validity of object tracking in laryngeal ultrasound video as a method to measure larynx height.

1 Introduction

Variation of larynx height (i.e., vertical larynx position) plays an important role in speech articulation and may act as a mechanism to sensitize formant frequencies to downstream area perturbations [1], to compensate for a declining subglottal pressure [2], and to control f_0 [3], for example. One possible reason that there are not many studies on larynx height could be the lack of a standard measurement method. A promising candidate for this is object tracking in laryngeal ultrasound videos, because in contrast to other measurement methods [4, 5, 6], it is easy-to-use, non-invasive, applicable to a wide variety of subjects, and capable of sufficient temporal resolution [7, 2, 8]. Besides that, ultrasound video is commonly used in research on speech articulation, e.g., to study velum movements [9] or tongue configuration [10]. Regarding larynx height in ultrasound video, one basic question that has not been answered yet is about the tracking accuracy under realistic conditions. Answering this could help to assess the general accuracy of the measurement method, and to find an optimum amongst different tracking algorithms or parametrizations.

To this end, Moisik et al. [2] measured the similarity between actual and estimated positions of a metal bar moving along a ruler in RGB video. One could argue that this test case is too abstract since, e.g., position was estimated in RGB video and not in ultrasound video. Moisik et al. [2] themselves underline that "ultrasound video is generally noisy and of reduced image clarity and consistency" with "various structures of the larynx continually moving into and out of the ultrasound beam's imaging plane during speech", which may well affect tracking accuracy but has not been included in its estimation yet. In the present study we adopted the original idea from Moisik et al. [2] to a more realistic test case using a larynx phantom. With this, we determined the optimal parameters for an established tracking algorithm based on optical flow [8]. We compared this algorithm to an algorithm based on template matching, which has fewer parameters and has not been used for object tracking in laryngeal ultrasound video yet.

2 Materials and Methods

2.1 Larynx phantom, instrumentation, and recording

A phantom that mimics the mechanical, sonographical and anatomical properties of the human larynx including the surrounding tissues was built using a pig larynx (see Fig. 1), as inspired by Schroeder et al. [11]. The larynx was procured from a local butcher and stored in a refridgerator for no longer than two days, during which the data was recorded. The surrounding tissues of the larynx like the extrinsic laryngeal muscles were still present, at least partially, but not the skin. There are several reviews on suitable phantom materials [12, 13], partly addressing physicians needing similar phantoms to practice ultrasound navigation during biopsy or intubation. In the present study, the skin was modeled by a pad of gelatine with admixed psyllium husk fibers, which were used to regulate the gelatine echogenicity [11, 14, 15]. The mixture of 167 ml water, 6.7 g gelatine, and 6.7 g fibres yielded a 1 cm thick pad. This thickness was needed to provide sufficient stability for the following experiments, and could be considered a worst case scenario in which ultrasound has to travel through much adipose tissue before reaching the larynx.

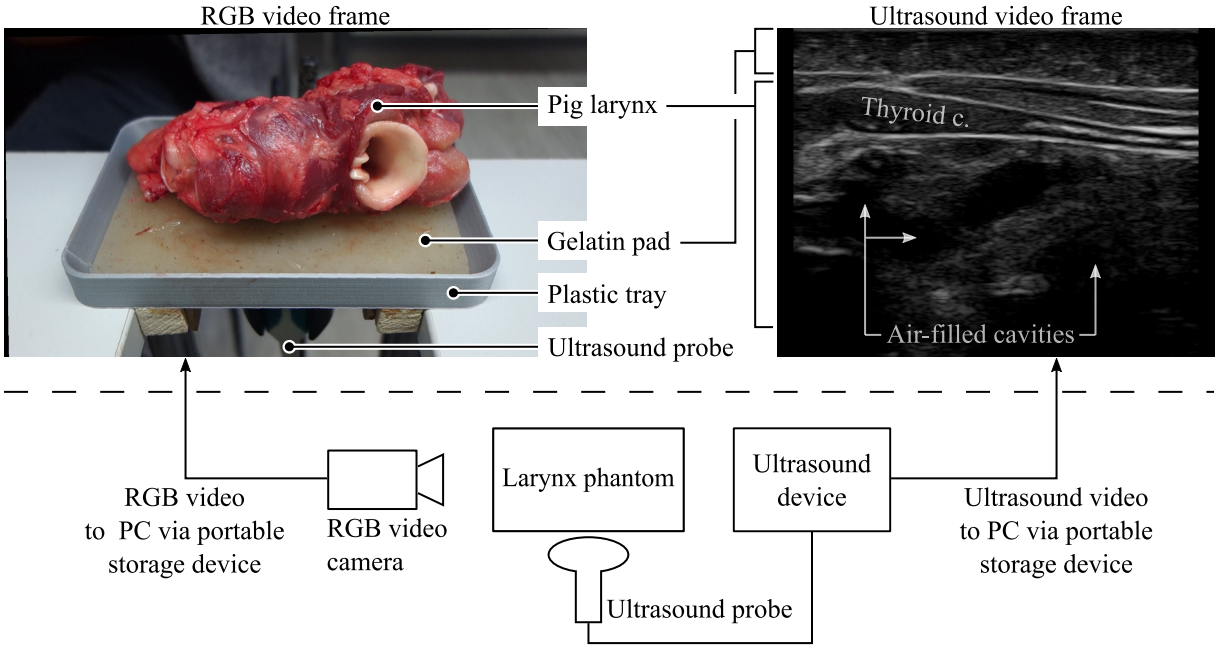

Figure 1 – Setup to record larynx phantom movements by means of RGB and ultrasound videos on a PC (area above the dashed line). Shown there are corresponding frames of both videos with the resting larynx. The RGB video frame illustrates the components of the larynx phantom including plastic tray, gelatine pad, and pig larynx. The ultrasound video frame illustrates the larynx anatomy including thyroid cartilage and air-filled cavities inside the larynx. Also visible are extrinsic laryngeal muscles between thyroid cartilage and gelatine pad, and vocal folds between tyroid cartilage and air-filled cavities.

A 3D-printed plastic tray that was fixed between two tabletops was holding the pad in position. The larynx, which could slide on the pad by means of a thin ultrasound gel film, was periodically moved from side to side by hand. A standard camera recorded the movement as RGB video with a frame rate of 25 Hz and a resolution of 1920×1080 pixels. At the same time, an ultrasound probe (SonoScape Phased Array 2P1) contacted the pad from below through a cutout in the plastic tray, to record video on an ultrasound device (SonoScape S2) in B-mode with a resolution of 640×480 pixels, a frame rate of 54 Hz, a probe pulse frequency of 10 MHz, and a scanning depth of 3.7 cm. For the further processing, each RGB video frame was rotated 0.9° counterclockwise to correct for a slight misalignment between camera and phantom.

In an *in vivo* measurement of larynx height, the ultrasound probe is typically placed on the skin in the thyroid lamina region [2, 19], and the larynx moves vertically. The larynx phantom models this, but with the movement axis aligned horizontally to keep the experimental setup as simple as possible. Therefore, horizontal larynx phantom position represents larynx height. Both the RGB and the ultrasound video are oriented in a way that rightward larynx phantom movement represents larynx raising and leftward movement larynx lowering.

2.2 Determination of larynx phantom position in RGB video

To determine the *actual* horizontal larynx phantom position in the RGB video as a reference for later, the custom Matlab R2021a GUI illustrated in Fig. 2 was used. The scrollbar controlled the frame index (black vertical line in Fig. 2c) and thereby the frame shown in Fig. 2a. In a first step, start and end of the recorded movement were determined roughly at frames 800 and 1300, respectively. In both frames and in a total number of 123 arbitrarily selected frames inbetween, a landmark was selected by means of a draggable marker (white circle in Fig. 2a). The landmark is detailed further below. The frames with selected landmarks are marked by the gray vertical lines in Fig. 2c. For all other frames, the landmark position was calculated using linear interpolation between the neighbouring selected landmark positions. The alignment between marker and landmark was checked visually across all frames. The horizontal larynx phantom position was determined as the horizontal marker position in each frame (black curve in Fig. 2c). The ruler in Fig. 2a served for conversion of the position from pixels to cm.

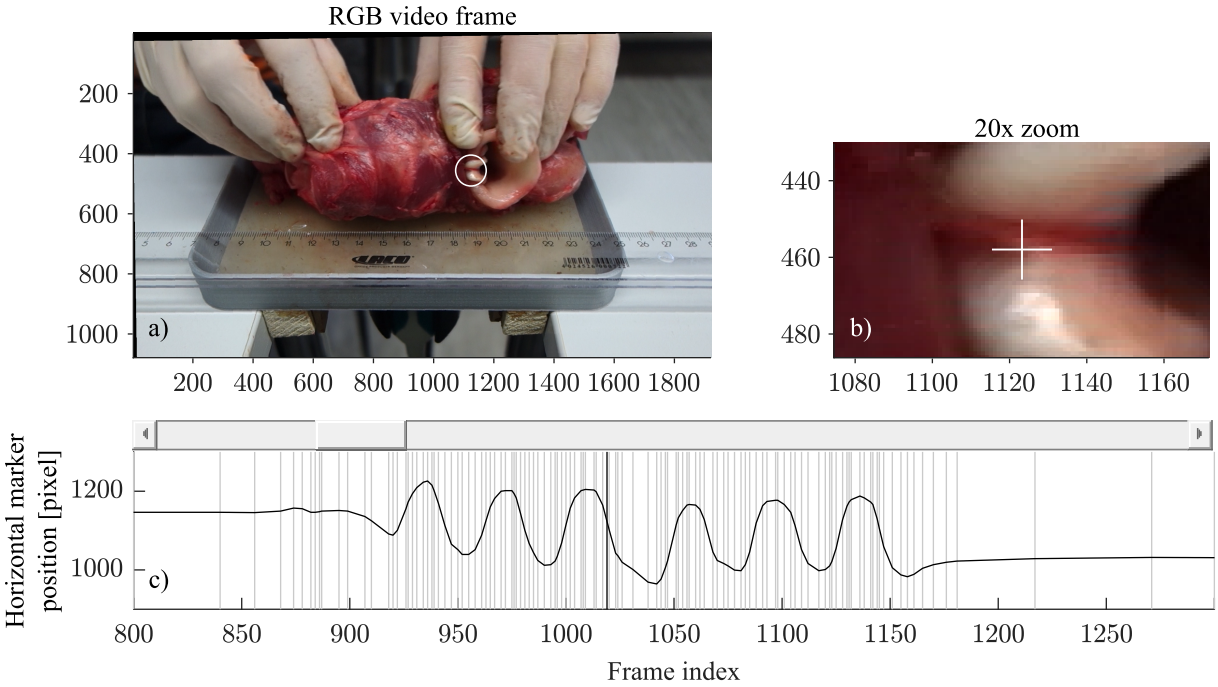

Figure 2 – Matlab GUI to determine the horizontal larynx phantom position in RGB video. Some GUI elements were rearranged for illustration purposes, or not shown like load and save buttons. a) Current RGB video frame with a white marker at the landmark. b) Twentyfold magnification of the marker region with a cross at the marker center. c) Horizontal marker position over time. The current frame, selected using the scrollbar, is shown as a black vertical line. Frames with a selected marker position are displayed as gray vertical lines. The linear interpolation between these frames is the *actual* horizontal larynx phantom position (black cuve).

The landmark mentioned above was defined at a groove in the cartilage structure exiting the pig larynx cranial (white circle in 2a). The groove appeared as a shaded triangle, the right corner

point of which was the landmark. What makes this a reasonable choice is that the cartilage structure seen in the RGB video is the apex of the so-called angle winding structure of the pig larynx [16], which directly connects to (and therefore should show identical movement as) the thyroid cartilage seen in the ultrasound video. For a more accurate placement of the marker, the marker region is shown in twentyfold magnification in Fig. 2b, where the cross is the marker center.

2.3 Estimation of larynx phantom position in ultrasound video

2.3.1 Algorithm based on optical flow

The horizontal larynx phantom position was estimated in ultrasound video using an established algorithm based on optical flow [8]. It first calculates displacement fields, also known as deformable grids, between adjacent frames, then calculates the weighted average of each displacement field to estimate larynx velocity, and finally integrates larynx velocity over time using cumulative trapezoidal numerical integration, similar to other established algorithms [7, 2]. The algorithm used here is available as the Matlab function `imregdemons()`, which computes each displacement field iteratively in a multi-scale scheme using image pyramids for execution speed and robustness, and Gaussian smoothing after each iteration for regularization [17]. The displacements fields were computed with seven parameter sets for the number of pyramid levels and iterations per pyramid level (see Tab. 1), and for three different standard deviations of the Gaussian smoothing kernel (0.5, 1, and 2 pixels), the combination of which led to 21 displacement field sequences. The Gaussian smoothing kernel was square-shaped with the edge

Table 1 – Seven parameter sets for the number of pyramid levels and iterations per pyramid level, where Sets 1 to 4 did not use the highest pyramid level. The lowest pyramid level corresponds to full image resolution. Incrementing it by one means reducing image resolution by a factor of four.

Pyramid level	**Number of Iterations**						
	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7
4 (coarse)	-	-	-	-	200	400	800
3	100	200	400	800	100	200	400
2	50	100	200	400	50	100	200
1 (fine)	25	50	100	200	25	50	100

length being an odd number of pixels of at least three standard deviations. Fig. 3a shows one displacement field in the region of visible homogenous movement. Especially the leftmost and rightmost 30 image pixels were excluded from this region because there, moving objects might not be contained in both adjacent frames and unwanted edge distortions of the displacement field are possible. In the shown region, horizontal displacements with an absolute value above a certain threshold were averaged to one estimate for horizontal larynx phantom velocity. The threshold was computed either as 0.5, 0.75, or 0.95 times the absolute horizontal displacement field maximum, which, via numerical integration, led to 63 (21 displacement field sequences $\times$ 3 thresholds) curves for horizontal larynx phantom position over time.

2.3.2 Algorithm based on template matching

The established algorithms based on optical flow describe the movement between adjacent frames in a relatively general way, but at the cost of increased algorithm complexity and number of parameters. Maybe, the movement could be described in a simpler way, namely as whole larynx shift in horizontal (and vertical) direction. A requirement for this would be a sufficiently

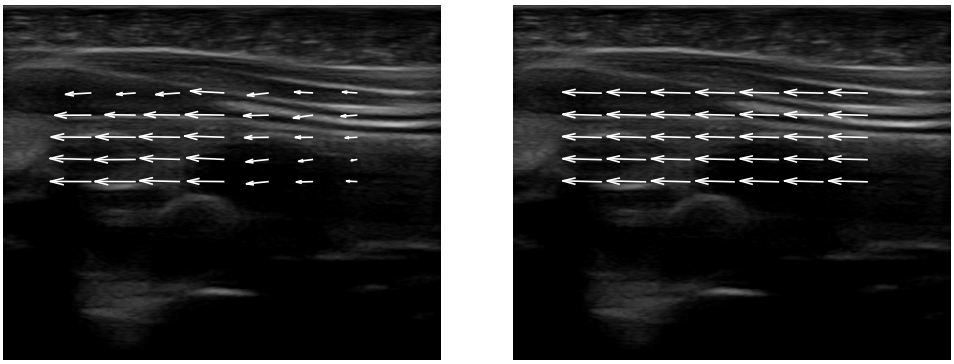

Figure 3 – Ultrasound video frame showing larynx movement from right to left, together with displacement fields in the region of visible homogenous movement, estimated using different algorithms. a) Algorithm based on optical flow. b) Algorithm based on template matching. For illustration purposes, both a) and b) do not show one displacement vector per pixel, as is actually the case, but only one displacement vector per 50×50 pixels, which was scaled up by a factor of three.

high frame rate, where other larynx movements due to, e.g., cricoarytenoid joint rotation, as well as artifacts due to movement into and out of the beam's imaging plane are small enough to be neglected. Based on these considerations, the following algorithm using template matching was considered as an alternative to the established algorithms. For each pair of adjacent frames, the template was defined by the region of visible homgenous movement (see Sec. 2.3.1) in the first frame. In the second frame, the template was shifted pixel-wise within a search region of ±30 pixels horizontally and ±2 pixels vertically around the original position. The distance between template and corresponding pixels of the second frame was computed as the sum of squared intensity differences. The best estimate for the shift, i.e., the one that leads to the minimal distance, was refined to subpixel accuracy using an algorithm that can be understood as average optical flow estimation in a single iteration at the lowest pyramid level without smoothing [18]. The resulting (single) displacement vector represents the average displacement of each pixel in the considered region, as illustrated in Fig. 3b. Horizontal larynx phantom position over time was calculated via numerical integration of horizontal vector components analogously to Sec. 2.3.1, but, in contrast to there, led to only one curve since no parameters were to be optimized here.

2.4 Analysis of tracking accuracy

In a first step, the reference curve for the horizontal larynx phantom position over time, as determined from RGB video in Sec. 2.2, was upsampled from 25 Hz to the ultrasound video frame rate of 54 Hz using linear interpolation. Then, the mean values were subtracted from the reference curve, and from the 63 estimated curves from Sec. 2.3.1 and the one from Sec. 2.3.2. In a next step, the cross-correlation functions between the reference curve and all estimated curves were calculated. Then, each estimated curve was shifted by the argument of its cross-correlation function maximum, to synchronize it with the reference curve. To measure tracking accuracy, Pearson's correlation coefficient between the reference curve and each (synchronized) estimated curve was calculated as the corresponding cross-correlation function maximum, divided by the number of samples and the standard deviations of the curves.

3 Results

Fig. 4 shows the reference curve (dashed curve) determined from RGB video, and two synchronized curves estimated from ultrasound video based on optical flow (black solid curve) and template matching (gray solid curve), respectively. By a visual assessment, the estimated curves

are very similar to each other and to the reference curve, except for a certain drift. After half of the tracking time, approximately, the estimated curves begin to drift away from each other, which is expressed by different tracking accuracies. Object tracking using template matching achieved 94 % accuracy, compared to which optical flow achieved a higher value of 97 % with the optimal parameter combination. The tracking accuracies for all 63 optical flow parameter combinations are shown in the matrix above the curves. The black field corresponds to the optimum, which was found at the smallest considered number of pyramid levels and iterations per pyramid level, the largest considered standard deviation for Gaussian smoothing, and the smallest considered displacement field thresholding factor. The gray fields correspond to tracking accuracies still higher than for template matching and the white fields to lower values.

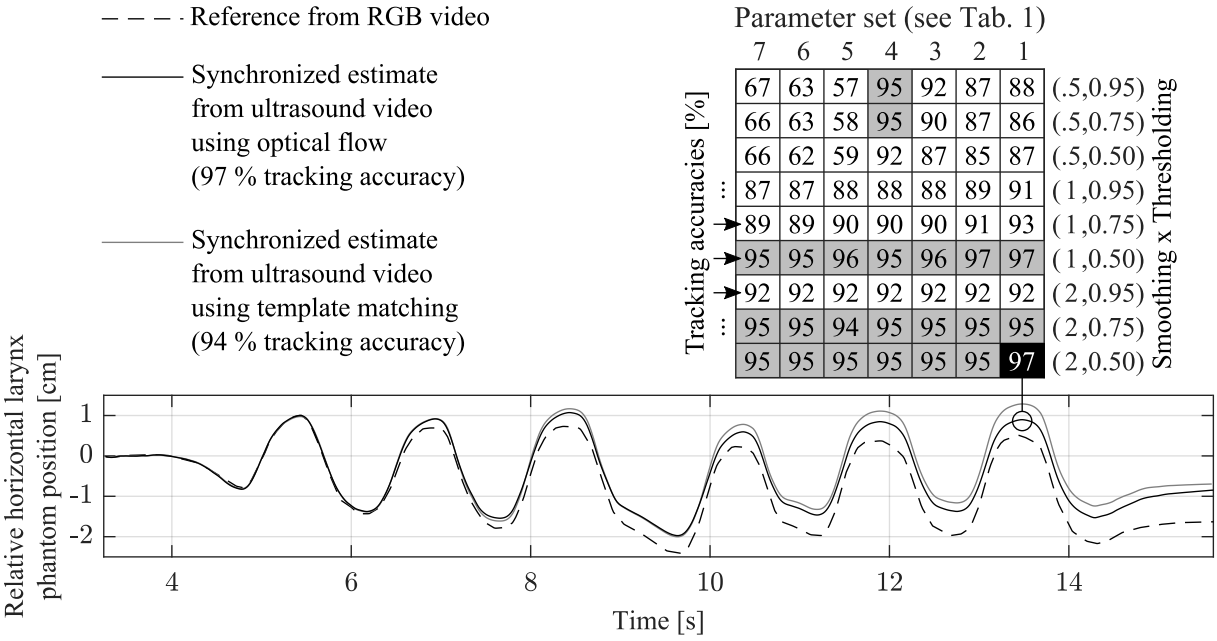

Figure 4 – Relative horizontal larynx positions over time, with the start positions shifted to 0 cm for illustration purposes. Shown are the reference curve (dashed curve), and the synchronized estimates from ultrasound video using optical flow and optimal parameters (black solid curve) and from template matching (gray solid curve). The top right matrix shows the tracking accuracies for all optical flow parameter combinations. Each column contains one parameter set for the number of pyramid levels and iterations per pyramid level (see Tab. 1), and each row contains one parameter combination for the standard deviation of the Gaussian smoothing kernel ("Smoothing"), measured in pixels, and the displacement field thresholding factor ("Thresholding"). Leading and trailing RGB video time intervals without substantial movement (see Fig. 2c) are not displayed here because they were not contained in ultrasound video and not considered in tracking accuracy measurement.

4 Discussion and Conclusion

The present study compared accuracies of two different object tracking algorithms based on optical flow and template matching, respectively, in laryngeal ultrasound videos using a phantom based on gelatine and an excised pig larynx. Both algorithms showed high accuracies of 97 % and 94 %, respectively, which supports the validity of object tracking in laryngeal ultrasound videos as a method to measure larynx height [7, 2, 8]. Since in the current study the tracking was done in the predefined region of visible homogenous movement, template matching could be implemented without any parameters to be specified by the user. Therefore, despite the lower achievable tracking accuracy, template matching may be preferred over optical flow, which used three parameters. One of them was the parameter set for the pyramid level and iterations per

pyramid level (see Tab. 1). It seemed as if the lowest considered values in parameter set 1 sufficiently resolved even large displacements, which could explain that the specific parameter set had only a small effect on tracking accuracy, compared to the other two optical flow parameters. These were the standard deviation of the Gaussian smoothing kernel and the displacement field thresholding factor. The top right matrix in Fig. 4 suggests that increased standard deviation and decreased thresholding factor both lead to increased tracking accuracy, possibly through decreased noise before and after displacement field averaging, respectively. Displacement field averaging was needed in optical flow based tracking to estimate the horizontal larynx phantom movement.

The usage of the larynx phantom was accompanied by some limitations. One limitation was that the phantom movement was generated externally by manually moving the larynx from side to side. This may lead to a more homogenous movement than *in vivo* vertical larynx movement, since effects like throid cartilage rotation and muscular thickening due to contraction were not modelled. To address this, recording ultrasound videos *in vivo* together with another established measurement method like thyroumbrometry [4] for reference could be a possible future direcion. Thyroumbrometry is a straight-forward measurement method for larynx height, but restricted to subjects with a marked thyroid protuberance. Another limitation was that we could not explain conclusively the drift between the reference curve and the estimated curves in Fig. 4. On the one hand, this could be due to a movement deviation in RGB video compared to ultrasound video, maybe because the angle winding structure deformes during phantom movement. On the other hand, this could be due to an accumulating tracking error. The fact that the drifts of the two estimated curves in Fig. 4 are in good accordance until half of the tracking time, approximately, and then begin to deviate from each other, suggests that both effects from above may be involved here. This limitation in the interpretability of the results is less critical for measurements of the vertical larynx phantom, where a possible drift is removed anyway [19]. In the future, more sophisticated trackers [20], which, among other advantages, are less prone to tracking error accumulation, could be included in the comparison of object tracking algorithms.

References

[1] WOOD, S.: *The acoustical significance of tongue, lip, and larynx maneuvers in rounded palatal vowels*. *The Journal of the Acoustical Society of America*, 80(2), pp. 391–401, 1986.

[2] MOISIK, S. R., H. LIN, and J. H. ESLING: *A study of laryngeal gestures in mandarin citation tones using simultaneous laryngoscopy and laryngeal ultrasound (sllus)*. *Journal of the International Phonetic Association*, 44(1), pp. 21–58, 2014.

[3] HONDA, K., H. HIRAI, S. MASAKI, and Y. SHIMADA: *Role of vertical larynx movement and cervical lordosis in f0 control*. *Language and Speech*, 42(4), pp. 401–411, 1999.

[4] EWAN, W. G. and R. KRONES: *Measuring larynx movement using the thyroumbrometer*. *Journal of Phonetics*, 2(4), pp. 327–335, 1974.

[5] GANDOUR, J. and I. MADDIESON: *Measuring larynx movement in standard thai using the cricothyrometer*. *Phonetica*, 33(4), pp. 241–267, 1976.

[6] KOB, M. and T. FRAUENRATH: *A system for parallel measurement of glottis opening and larynx position*. In C. MANFREDI (ed.), *Fifth International Workshop on Models*

and Analysis of Vocal Emissions for Biomedical Applications, Firenze, Italy, pp. 109–111. 2007.

[7] MOISIK, S. R.: *The epilarynx in speech*. Ph.D. thesis, University of Victoria, 2013.

[8] YUN, D. P. Z. and S. R. MOISIK: *A laryngeal ultrasound study of singaporean mandarin tones*. In S. CALHOUN, P. ESCUDERO, M. TABAIN, and P. WARREN (eds.), *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia*. 2019.

[9] BIRKHOLZ, P. and C. KLEINER: *Velocity differences between velum raising and lowering movements*. In A. KARPOV and R. POTAPOVA (eds.), *International Conference on Speech and Computer, St. Petersburg, Russia*, pp. 70–80. 2021.

[10] LAWSON, E., J. M. SCOBBIE, and J. STUART-SMITH: *Bunched /r/promotes vowel merger to schwar: An ultrasound tongue imaging study of scottish sociophonetic variation*. *Journal of Phonetics*, 41(3-4), pp. 198–210, 2013.

[11] SCHROEDER, K. M., J. RAMAMOORTHY, and R. E. GALGON: *An easily made, low-cost phantom for ultrasound airway exam training and assessment*. *Indian Journal of Anaesthesia*, 57(1), p. 31, 2013.

[12] CULJAT, M. O., D. GOLDENBERG, P. TEWARI, and R. S. SINGH: *A review of tissue substitutes for ultrasound imaging*. *Ultrasound in Medicine & Biology*, 36(6), pp. 861–873, 2010.

[13] ZELL, K., J. I. SPERL, M. W. VOGEL, R. NIESSNER, and C. HAISCH: *Acoustical properties of selected tissue phantom materials for ultrasound imaging*. *Physics in Medicine & Biology*, 52(20), p. N475, 2007.

[14] RICHARDSON, C., S. BERNARD, and V. A. DINH: *A cost-effective, gelatin-based phantom model for learning ultrasound-guided fine-needle aspiration procedures of the head and neck*. *Journal of Ultrasound in Medicine*, 34(8), pp. 1479–1484, 2015.

[15] SEGUIN, J. and M. O. TESSARO: *A simple, inexpensive phantom model for intubation ultrasonography training*. *Chest*, 151(5), pp. 1194–1196, 2017.

[16] GAO, N., X. CUI, G. SUN, G. ZHANG, G. ZHOU, and X. ZHAO: *Comparative anatomy of pig arytenoid cartilage and human arytenoid cartilage*. *Journal of Voice*, 33(5), pp. 620–626, 2019.

[17] THIRION, J.-P.: *Image matching as a diffusion process: an analogy with maxwell's demons*. *Medical image analysis*, 2(3), pp. 243–260, 1998.

[18] CHAN, S. H., D. T. VÕ, and T. Q. NGUYEN: *Subpixel motion estimation without interpolation*. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 722–725. 2010.

[19] KLEINER, C., P. HÄSNER, and P. BIRKHOLZ: *Intrinsic velocity differences between larynx raising and larynx lowering*. *[Manuscript submitted for publication]*, 2023.

[20] LUKEŽIČ, A., T. VOJÍŘ, L. ČEHOVIN ZAJC, J. MATAS, and M. KRISTAN: *Discriminative correlation filter with channel and spatial reliability*. In *Proceedings of the IEEE conference on computer vision and pattern recognition, Honolulu, Hawaii*, pp. 6309–6318. 2017.