AUTOMATIC PALATE DELINEATION IN ULTRASOUND VIDEOS

Guillaume Faucher[1], Elham Karimi[1], Lucie Ménard[2] and Catherine Laporte[1]

[1]École de technologie supérieure; [2]University of Québec in Montréal
guillaume.faucher.1@etsmtl.net, elham.karimi.1@etsmtl.net, menard.lucie@uqam.ca, catherine.laporte@etsmtl.ca

ABSTRACT

Measurements of palate location can assist ultrasound (US)-based analysis of articulatory tongue motion by providing complementary information about oral cavity constriction. They also provide a rigid reference frame relative to which tongue measurements from different time points can be registered. Locating the palate in US images is challenging because it is generally invisible except during swallowing, and even then, it is often not readily recognizable in any single frame. This paper introduces a new automated method to extract a palate contour from an US video acquired during swallowing. The method is based on a *cumulative echo skeleton image*, which highlights structures that are consistently located over time. In experiments with 22 US videos, most of the automatically extracted palate traces were within 3 mm of a manual palate trace in terms of mean sum of distances error, demonstrating the potential of the proposed approach.

Keywords: Ultrasound, palate, image processing

1. INTRODUCTION

Ultrasound (US) imaging is ideally suited and widely used to study tongue shape and motion during speech. However, tongue location and shape alone fail to fully capture phonetically relevant variables pertaining to constrictions of the oral cavity. To remediate this, it is useful to measure the configuration of the tongue *relative to the palate*, which in turn requires locating the palate in the same reference frame as the tongue. Palate measurements also provide a rigid reference frame for registration of US tongue data over time [9], and for studying the influence of the palate shape on articulation [2].

While dental casts provide detailed palate measurements (e.g., in electropalatography), measuring the palate from the US images themselves is quick and inexpensive in comparison [12]. Unfortunately, the palate is usually invisible in US images because of the air separating it from the tongue. Recently, Wrench [14] demonstrated an indirect, "vocal tract carving" approach that automatically infers the palate surface as the upper boundary of the space reached by the (automatically tracked) tongue as speech is continuously elicited from a speaker. This method is appealing because it can be used in real-time. However, it requires high speed imaging to capture the instants where the tongue touches the palate. It also requires eliciting such contacts over the entire oral cavity, which can be challenging in populations with speech impairments.

The traditional approach is to directly delineate the palate in recordings where the speaker is swallowing or holding liquid in his/her mouth [13, 3, 9]. Then, US can reach the palate, causing a visible echo in the US images. Edgetrak [8] offers a rudimentary semi-automatic tool to fit a snake to manually annotated palate points on a single image. However, in many cases, the palate is only partially visible in any given image. Thus, Epstein and Stone [3] recommend using a short video clip acquired during swallowing and manually accumulating partial palate traces over a series of frames as a bolus of water or saliva travels through the mouth and different parts of the palate become visible. This is challenging for the operator, who must simultaneously see through time and space to delineate the palate in a piecewise fashion from one frame to the next.

This paper proposes a new automatic method to extract the mid-sagittal palate contour from US videos of a swallow that overcomes the aforementioned difficulty. The method, described in Section 2, is based on a *cumulative echo skeleton image*, which (1) highlights features of the image that are consistently located over time (and might correspond to the palate) and (2) connects the parts of the palate that are visible in different frames as a single structure. Section 3 describes experiments on 22 swallow videos, wherein automatically extracted palatal traces were compared against a manual trace, with promising results. A discussion, including directions for future work, is presented in Section 4.

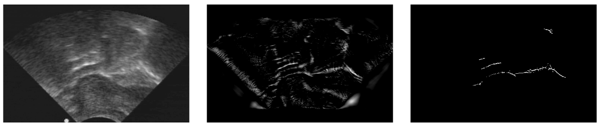
2. METHOD

The proposed palate extraction method comprises processing at the level of (1) the individual frames composing the swallow video sequence and (2) the sequence itself. The individual image processing step, described in Section 2.1, extracts a line drawing, called a *skeleton*, that characterizes the shape of the brightest ridge-like structures (echoes) in the image (Fig. 1, right). At the sequence level, the skeletons from the individual images are summed over time, leading to a *cumulative echo skeleton image* (Fig. 2), which emphasizes the structures that are most persistent and consistently located over the duration of the video sequence. Since the palate is mostly rigid and immobile (unlike the tongue or imaging artefacts), the cumulative echo skeleton image carries meaningful information about its shape and location. This information is extracted using thresholding, clustering and robust curve fitting operations prior to shape refinement using a snake fitted to one or more of the original US images, as shown in Fig. 3 and detailed in Section 2.2.

2.1. Frame-level processing

Fig. 1 shows the processing steps applied to each image in the swallow sequence. The bright ridge-like echoes in the US image, typically the palate and tongue, along with some imaging artefacts, are enhanced within a phase symmetry map [7, 6]. This map is obtained by first filtering the image using odd and even log-Gabor filters at different scales and orientations (5 scales and 14 orientations were used here). Phase symmetry is the average amplitude difference in the responses from the even and odd filters or, more intuitively, the degree of even symmetry or "ridgeness" of the structures in the image. Fig. 1 (middle) shows a typical result.

Figure 1: Processing of individual images. From left to right: original image, phase symmetry map, skeleton of thresholded phase symmetry image.



The phase symmetry map is thresholded to preserve only the brightest and largest echoes from the original images. A *skeleton* of these structures then extracted by finding their *medial axis*. The medial axis is the locus of points within a shape that are equidistant from two or more of the shape boundary points [1]. Many methods exist to compute medial axes; here, Rezanejad et al's robust average outward flux method [11] is used. Typical skeletons are shown in Figs. 1 (right) and 2.

2.2. Sequence level processing

The echo skeletons computed from individual US images typically contain information about parts of the tongue and/or palate and/or some imaging artefacts. One skeleton is generally insufficient to infer the palate surface in a robust manner. For this, one must exploit the temporal information contained in the sequence of images. Thus, the skeletons extracted from the different images in the sequence are summed to form a *cumulative echo skeleton image*, as shown in Fig. 2. In this sum, each white pixel from an individual skeleton image slightly increases the intensity of the corresponding pixel in the cumulative echo skeleton image. Immobile and persistent structures like the palate contribute to similar locations in the cumulative echo skeleton image over time, leading to high signal levels, whereas moving or non-persistent structures like the tongue or imaging artefacts, though often brighter than the palate in single images, contribute to more diverse locations and lead to weaker signals.

Fig. 3 shows how the cumulative echo skeleton image is processed to extract palate contours. Otsu thresholding [10] is applied to its non-zero pixel intensities to remove noise arising from non-persistent structures in the US images. The locations of the non-zero valued pixels in the thresholded image are clustered using DBSCAN [4], an algorithm that forms arbitrary numbers of clusters from spatial data based on thresholds ε, the maximum distance between points within a cluster, and *MinPts*, the minimum acceptable number of points within each cluster. Here, ε = 20 pixels and MinPts = 10. The widest cluster is selected as potentially containing the palate. Within this cluster, the point of maximal height is retained for each position along the horizontal axis. This favours points arising from the reflection of US off the palate rather than off the tongue. A second order polynomial is then fitted to the resulting points using RANSAC [5], a robust fitting algorithm that finds the least-squares fitting curve accounting for the largest number of inliers, while rejecting outliers. Outliers are defined by a maximum allowable distance to the curve (10 pixels in this work). A cubic spline is then fit to the inliers, which is then used to initialize a snake fit [8] to the palate for refinement.

3. EXPERIMENTAL RESULTS

Automatic palate extraction was tested on US data from 6 healthy subjects, 3 adults (A1-A3) and 3 chil-

Figure 2: Creation of the cumulative echo skeleton image. The 6 leftmost panels show sample US images from a swallowing video sequence (top) and the skeletons extracted from each one (bottom). The rightmost panel shows the cumulative echo skeleton image computed from the sum of the skeletons over time.

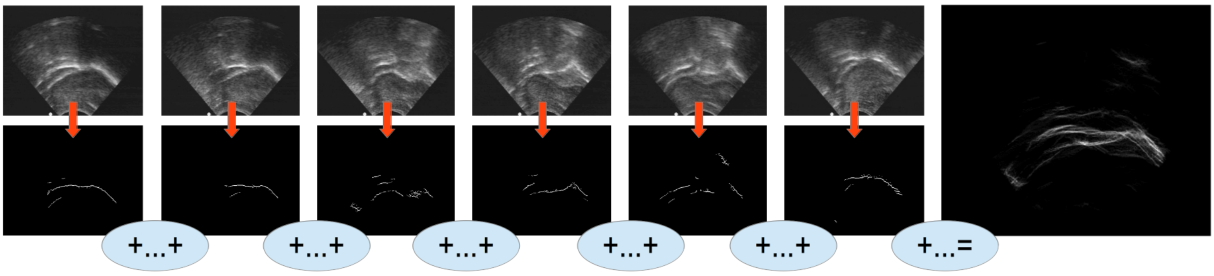
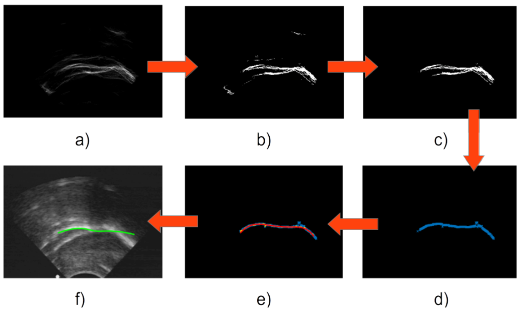


Figure 3: Palate contour extraction from the cumulative echo skeleton image. The cumulative echo sekeleton image (a) is thresholded (b) to enhance temporally persistent structures. The widest cluster of white pixels is extracted from the thresholded image (c). Data from the earliest US echoes are removed (d), and a cubic spline is robustly fitted to the remaining points (e). A snake is fitted to US images to obtain palate contours (f).



dren (C1-C3), acquired during speech. Twenty-two clips with swallowing, ranging in length from 48 to 154 frames, were manually extracted from the full recordings. A reference palate contour was manually traced on one reference image in each clip. The mean sum of distances (MSD) between the reference palate trace u and the automatic palate trace v was computed as

$$\text{(1)}\ \ \mathrm{MSD}(u,v) = \frac{1}{m+n}\Big(\sum_{i=1}^{n}\min_{j}||\boldsymbol{v_i}-\boldsymbol{u_j}||+\sum_{j=1}^{m}\min_{i}||\boldsymbol{u_j}-\boldsymbol{v_i}||\Big),$$

where $\boldsymbol{u_i}$ (respectively $\boldsymbol{v_j}$) is the vector of x and y coordinates of the ith (respectively jth) vertex of u (respectively v), $i \in \{1,\ldots,m\}$ and $j \in \{1,\ldots,n\}$.

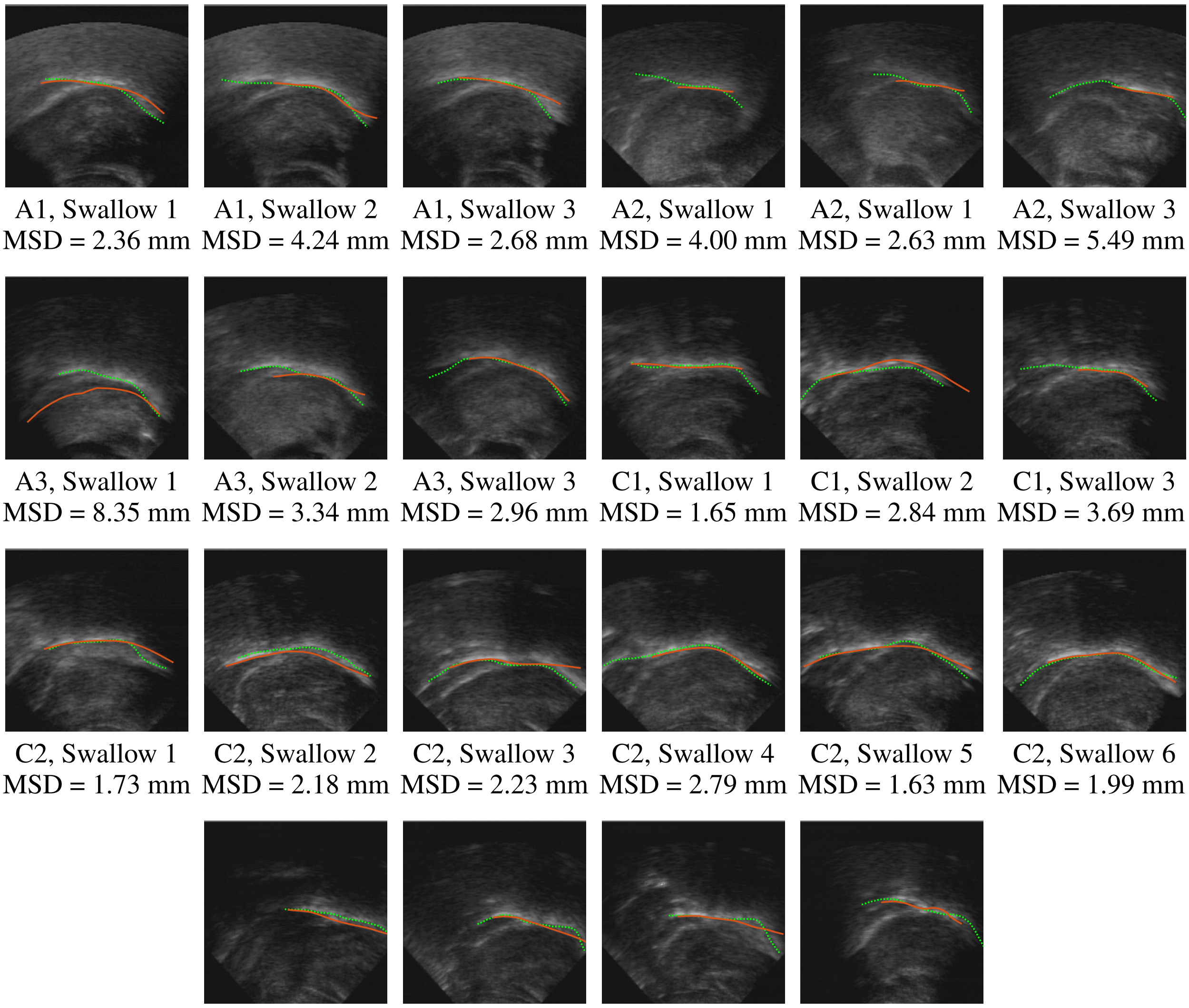
Fig. 4 compares the automatically extracted palate traces with the reference palate traces and reports the MSD between them. Generally, the automatically detected palate traces overlap fairly well with the manual ones. However, the automatically detected traces tend to be shorter, particularly towards the back of the mouth. In several clips from child subjects, the automatic palate trace matches quite well to the location of the hard palate, which is rigid and generally more visible in the images, whereas the manual trace often also comprises the velum. This suggests that the proposed method is most successful at locating the rigid part of the palate. Arguably, this is desirable for applications requiring measurements of the relative configuration of the tongue with respect to a rigid palate reference frame.

The method failed in a few cases. For subject A2, it only detected a small segment of the palate due to the relatively poor quality of the images in this subject's recordings, where even the tongue was less visible than in other recordings. In subject A3, the method detected the tongue instead of the palate in two of three swallowing clips. Upon inspection, both clips were found to depict a resting, fairly immobile tongue, for many frames before and after the swallowing motion. Thus, the cumulative echo skeleton image contained stronger contributions from the resting tongue than from the palate. This points to the importance of feeding quality input to the method. Ideally, this would include little but the actual swallowing motion.

4. CONCLUSIONS

This paper presented a new method to automatically extract the mid-sagittal palate contour from US video sequences of swallowing by exploiting the

Figure 4: Automatic palate extraction results (solid red) and manual palate trace (dashed green) overlayed on the reference image from each swallowing sequence. The MSD between the two traces is reported below each image.



A1, Swallow 1 MSD = 2.36 mm

A1, Swallow 2 MSD = 4.24 mm

A1, Swallow 3 MSD = 2.68 mm

A2, Swallow 1 MSD = 4.00 mm

A2, Swallow 1 MSD = 2.63 mm

A2, Swallow 3 MSD = 5.49 mm

A3, Swallow 1 MSD = 8.35 mm

A3, Swallow 2 MSD = 3.34 mm

A3, Swallow 3 MSD = 2.96 mm

C1, Swallow 1 MSD = 1.65 mm

C1, Swallow 2 MSD = 2.84 mm

C1, Swallow 3 MSD = 3.69 mm

C2, Swallow 1 MSD = 1.73 mm

C2, Swallow 2 MSD = 2.18 mm

C2, Swallow 3 MSD = 2.23 mm

C2, Swallow 4 MSD = 2.79 mm

C2, Swallow 5 MSD = 1.63 mm

C2, Swallow 6 MSD = 1.99 mm

C3, Swallow 1 MSD = 1.63 mm

C3, Swallow 2 MSD = 1.46 mm

C3, Swallow 3 MSD = 1.81 mm

C3, Swallow 4 MSD = 2.38 mm

persistence of the echoes generated by the palate over time. The method was tested on 22 video sequences with promising results in terms of accuracy. In future work, ideal or near ideal swallowing sequences could probably be extracted automatically from larger speech video recordings by searching for segments with large amounts of motion and weak acoustic signal. Automated palate extraction using spontaneous swallowing in US video sequences is an important step towards facilitating more meaningful articulatory measurements. It may also provide a useful rigid reference frame which can help evaluate and compensate for some types of head motion (e.g. front to back) in the analysis of US recordings in the field, where sophisticated head motion measurement devices may not be practical or available.

ACKNOWLEDGMENTS

This work was supported by the Natural Sciences and Engineering Research Council of Canada, the Social Science and Humanities Research Council of Canada and the Fonds de Recherche Québécois - Nature et Technologie.

5. REFERENCES

[1] Blum, H. 1967. A transformation for extracting new descriptors of shape. *Models for the Perception of Speech and Visual Form* (5), 362–380.

[2] Brunner, J., Fuchs, S., Perrier, P. 2009. On the relationship between palate shape and articulatory behavior. *Journal of the Acoustical Society of America* 125(6), 3936–3949.

[3] Epstein, M., Stone, M. 2005. The tongue stops here: ultrasound imaging of the palate. *Journal of the Acoustical Society of America* 118(4), 2128–2131.

[4] Ester, M., Kriegel, H. P., Sander, J., Xu, X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. *ACM SIGKDD Conference on Knowledge Discovery and Data Mining* 226–231.

[5] Fischler, M. A., Bolles, R. C. 1981. Random sample consensus: a paradigm for model fitting with applications of image analysis and automated cartography. *Communications of the ACM* 24(6), 381–395.

[6] Hacihaliloglu, I., Abugharbieh, R., Hodgson, A. J., Rohling, R. N. 2009. Bone surface localization in ultrasound using image phase-based features. *Ultrasound in Medicine & Biology* 35(9), 1475–1487.

[7] Kovesi, P. 1999. Image features from phase congruency. *Videre: Journal of Computer Vision Research* 1(3), 1–26.

[8] Li, M., Kambhamettu, C., Stone, M. 2005. Automatic contour tracking in ultrasound images. *Clinical Linguistics and Phonetics* 19(6-7), 545–554.

[9] Mielke, J., Baker, A., Archangeli, D., Racy, S. 2005. Palatron: a technique for aligning ultrasound images of the tongue and palate. *Coyote Papers: Working Papers in Linguistics, Linguistic Theory at the University of Arizona* 14, 96–107.

[10] Otsu, N. 1979. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man and Cybernetics* 9(1), 62–66.

[11] Rezanejad, M., Siddiqi, K. 2013. Flux graphs for 2d shape analysis. *Shape Perception in Human and Computer Vision* 41–54.

[12] Scobbie, J. M., Stuart-Smith, J., Lawson, E. 2008. Looking variation and change in the mouth: developing the sociolinguistic potential of ultrasound tongue imaging. Technical report Queen Margaret University.

[13] Stone, M. 2005. A guide to analysing tongue motion from ultrasound images. *Clinical Linguistics and Phonetics* 19(6-7), 455–501.

[14] Wrench, A. A. 2017. Real-time tongue contour fitting and vocal tract carving. *Ultrafest VIII* 99–100.