Automatic Motion Analysis of the Tongue Surface from Ultrasound Image Sequences

Yusuf Sinan Akgul   Chandra Kambhamettu
Department of Computer and Information Sciences
University of Delaware
Newark, Delaware 19716
{akgul,chandra}@cis.udel.edu

Maureen Stone
Division of Otolaryngology
The University of Maryland Medical School
Baltimore, Maryland 21201
mstone@surgery2.ab.umd.edu

Abstract

We present a system for automatic 2D analysis of the tongue movement from digital ultrasound image sequences. The system focuses on extraction, tracking and analysis of the tongue surface during speech production and swallowing. The input to the system is provided by a Head and Transducer Support System (HATS), which is developed for use in ultrasound imaging of tongue movement. We developed a novel active contour (snakes) model that uses several adjacent images during the extraction of the tongue surface contour for an image frame. The user supplies an initial contour model for a single image frame in the sequence. This initial contour is a form of expert knowledge input to the system, which is used to find the candidate contour points in the adjacent images. Subsequently, the new snake mechanism is applied to estimate optimal contours for each image frame using these candidate points. Finally, the system uses a postprocessing method to refine the positions of the contours by utilizing more spatiotemporal information.

We extended our previous work by applying the system to different speech and swallowing sequences using various constraints. The extended system can also extract qualitative local deformations with only a minimal computational overhead, which may be useful for the diagnosis of Cerebellar Ataxic disorder. We tested the system on several different speech and swallowing sequences produced by HATS. During the tests, we saw that the system is flexible enough to be used in a wide variety of cases. In addition, visual inspection of the detected contours by the speech experts confirms that the results are very promising and this system can be effectively employed in speech and swallowing research.

1. Introduction

In medical imaging, automatic extraction and tracking of the tongue surface can provide valuable information for speech and swallowing research. It has several application areas, including disordered speech, aging speech, linguistics, speech processing, and tongue modeling. In addition, a system developed for this purpose can also be applied to other medical edge-detection research such as heart and fetal measurements.

Ultrasound imaging is one of the most attractive ways of taking a sequence of pictures of the tongue during speech or swallowing because it can produce real-time capture rates and it is non-invasive. Alternative methods are either too slow to record movement, such as MRI, or they expose subjects to radiation, like X-rays.

In order to automatically extract and detect the tongue surface from an image sequence, we developed a system that uses ultrasound images of the tongue produced by HATS[8]. Although HATS solved several difficulties in taking reliable and accurate ultrasound tongue images, there are problems that should still be addressed(Figure 1). Some of these problems are unique to the ultrasound imaging and the imaging of the tongue movement. First, the ultrasound images are quite noisy. Second, structures within the tongue (tendons, blood vessels) and noise echo artifacts cause high contrast edges unrelated to the structure of interest. In addition, the ultrasound transducer of HATS[8] emits a beam upward from the chin. During the production of some speech sounds, portions of the tongue may be almost parallel to the ultrasound waves. Therefore, the edge corresponding to the surface of interest can be interrupted in places where the tongue does not reflect the waves. More discontinuity is caused when a contact occurs between the tongue and the palate or the teeth, which prevents the ultrasound waves from reflecting off the surface of the tongue. Another interesting problem arises during swallowing when

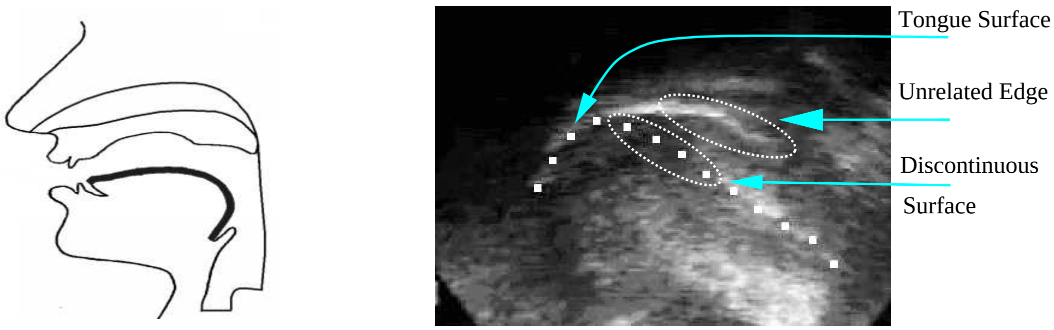


Figure 1. ***Left*-Position of the tongue surface inside the vocal tract. *Right*-Problems with the ultrasound images from HATS. (The front of the tongue is on the left. This image is taken from a swallowing sequence.)**

the swallowed material creates a disturbance for the ultrasound waves. Finally, even for human speech experts, it is necessary to look at the image sequence to be able to extract the tongue surface for a single poor image frame. As a result, standard ways of detecting edges do not work for this problem, and hence specific processing methods should be developed.

Our system uses a deformable model based on active contour models[5] as the main tool for detecting the contour of the tongue surface. Considering the problems listed above, snakes are attractive. First, snakes facilitate initial contours, so that an initial contour model can be given to the system by the user. This will help the system ignore unrelated high-contrast edges that are not similar to the contour model. The initial model can also be used to provide expert knowledge to the system. Second, snakes can enforce continuity and smoothness, which will help the system to estimate edge positions even in places where the surface of interest is interrupted. Interrupted edges are one of the most challenging problems of the tongue surface extraction process. Due to the physically-based nature of the active contour models, they can fit acceptable splines to the regions where the surface is interrupted. Finally, our experience shows that the snake mechanism can easily be modified to reflect some of the specific constraints of the problem, e.g., enforcing an angle range for the surface contour of some portion of the tongue.

A carefully chosen image frame and an initial model by a speech expert is crucial for the system to find the desired contours. Detecting the tongue surface of a given ultrasound image is a very difficult task even for a human if he or she is not trained. Therefore, a system to analyze ultrasound tongue images should be very flexible in order to take advantage of speech expert input. This extension of our previous work[1] was developed to consider this requirement. We were able to analyze tongue movement during swallowing by taking advantage of the flexibility of our system.

2. Deformable model

In this section, we will first explain the deformable model used for a single image frame, which uses a novel extension of active contour models. Later, we will explain how this model is used to process a complete sequence of ultrasound images.

2.1. Formulation for a single frame

Our formulation of active contour models is based on the discrete version of the original formulation[5] with some modifications. A snake is an ordered set of points $V = [v_1, v_2, ..., v_n]$ with a snake energy value calculated using internal and external forces. The internal forces serve to impose smoothness on the contour and similarity to the model contour, and external forces push the snake toward salient image features, i.e., edges. Given an initial model contour $S = [s_1, s_2, ..., s_n]$, we write the total energy of a snake, V, on image frame I as follows:

$$E_{Snake}(V, S, I) = \sum_{i=1}^{n} \alpha E_{int}(v_i, S) + \beta E_{ext}(v_i, I). \quad (1)$$

$E_{int}(v_i, S)$ is the internal energy of v_i which is the weighted sum of smoothness, $E_{smo}(v_i)$, and similarity to the initial model, $E_{sim}(v_i, S)$:

$$E_{int}(v_i, S) = \lambda E_{smo}(v_i) + (1 - \lambda) E_{sim}(v_i, S) \quad (2)$$

where α, β and λ are the weighting parameters.

$E_{smo}(v_i)$ is analogous to the smoothness term in the original snake formulation[5]. It is estimated as $E_{smo}(v_i) = 1 - \cos\theta_i$, where $i = 2 \ldots (n-1)$ and θ_i is the angle between $\overrightarrow{v_{i-1}v_i}$ and $\overrightarrow{v_i v_{i+1}}$. We define $E_{smo}(v_1) = E_{smo}(v_2)$ and $E_{smo}(v_n) = E_{smo}(v_{n-1})$.

Given the formula for $E_{smo}(v_i)$ and if $\alpha = 1, \beta = 0$ and $\lambda = 1$, the snake minimization will produce a straight line due to the missing external or similarity forces.

If it is used as the only internal energy term, this smoothness term suffers from some of the problems with the original smoothness constraints[5]. For example, it tends to shrink around strong edge points because closer the snake element positions on a contour, smaller the value of θ. However, this does not create a problem for our system because the internal energy of the snake uses this smoothness constraint in combination with the similarity constraint. If two snake elements get very close due to a strong edge, the internal energy will increase due to the dissimilarity to the model.

The main function of the similarity term is to propagate contour information provided by the speech expert to the surface extraction processes of all image frames. The information provided by the speech expert is properly utilized in this way. For each v_i, two similarity parameters, t_i and u_i, are extracted by $s_i = t_i s_{i-1} + u_i s_{i+1}$, where s_{i-1}, s_i, and s_{i+1} are the elements of the contour model S. These two parameters are then used to calculate the similarity of the position of v_i to the position of s_i with respect to neighboring elements. Figure 2 shows the calculation of similarity term for v_i using already extracted parameters t_i and u_i.

$$E_{sim}(v_i) = \frac{1}{l(V)} \left| v_i - (t_i v_{i-1} + u_i v_{i+1}) \right|^2$$

where $l(V)$ is the average distance between elements of V.

$$l(V) = \frac{1}{n-1} \sum_{i=2}^{n} \left| v_i - v_{i-1} \right|^2 .$$

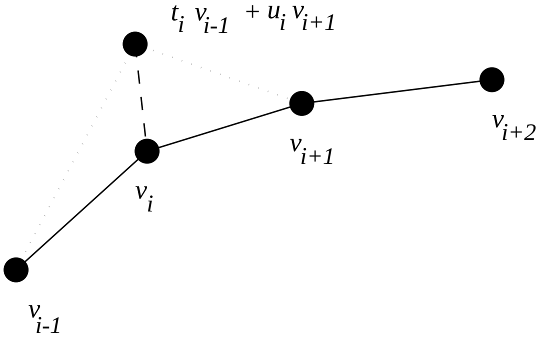


Figure 2. Similarity term for v_i is the length of the dashed line divided by $l(V)$

The similarity term $E_{sim}(v_i, S)$ is similar to the one proposed by Lai and Chin[6]. However, it is employed in a completely different way. Lai and Chin[6] use this energy as the only internal energy for their deformable model. In our system, it is uniquely used in combination with smoothness term.

The energy produced by the similarity term has another function. It can be used to qualitatively measure the deformation occurring around a snake element. Although, the energy values do not represent the local strain or they do not have meaningful units, it is useful to know what part of tongue deforms most for some applications. Section 3 explains how we use similarity term to extract qualitative local deformation information.

The external energy term $E_{ext}(v_i, I)$ is given by the negative of the image gradient at v_i

$$E_{ext}(v_i, I) = - \left| \nabla I(v_i) \right| .$$

There have been a number of proposed methods to minimize the Equation (1). Among these, we used the one proposed by Amini et. al.[2], which uses dynamic programming for solving variational problems in vision. This method takes advantage of locality nature of the energy calculations to achieve an exhaustive minimization in polynomial time. We modified this method such a way that only a useful subset of intensity points are fed to the minimization process[1]. With this method, in addition to the reduction in the computation time, it is possible to impose additional constraints on the relations between the snake elements. For example, our system imposes different angle-range constraints for different parts of the tongue surface by utilizing hard constraints instead of adding new terms to the energy function. Hard constraints are better choice for angle restrictions because we know that the tongue surface cannot have some angles at some positions. Using these constraints, some impossible element positions are automatically rejected before entering the minimization process, which reduces the cost of minimization and increases the robustness of the system.

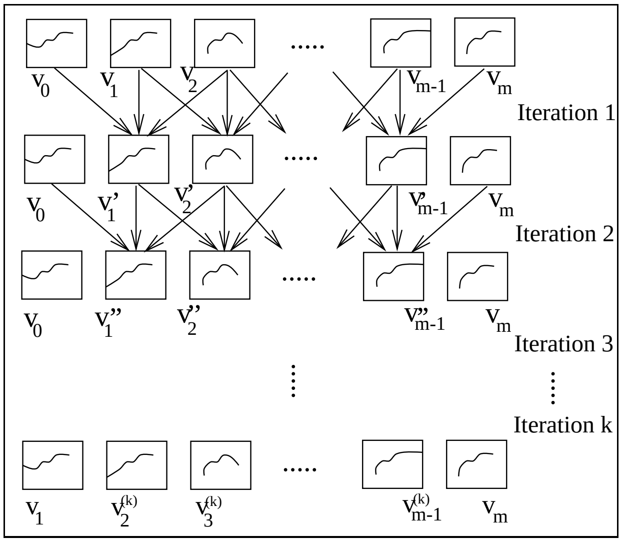
2.2. Formulation for the sequence

The formulations listed above would be sufficient if the problem were to find the contour for just one image frame. However, our problem requires extracting a set of contours from an image sequence. We used the following formula to minimize the energies of a set of snakes for the image sequence.

$$E_{Sequence} = \sum_{i=1}^{m} \min E_{Snake}(V_i, V_{i-1}, I_i) \qquad (3)$$

where V_0 is the initial contour given by the user.

Equation (3) implies that only the model snake is used in the current frame's contour extraction as the information from the image sequence. This may limit the system's capabilities because as described in the problem description, the system should use as much information as possible from the whole sequence. In order to overcome this limitation,



```
$Energy1 = BIG - NUMBER$
while($\sum_{i=1}^{m} E_{Snake}(V_i, V_{i-1}, I_i) < Energy1$) do
    $Energy1 = \sum_{i=1}^{m} E_{Snake}(V_i, V_{i-1}, I_i)$
    for($i = 1; i < (m-1); i = i+1$)
        Use the postprocessing algorithm
        to refine the position of $V_i$,
        using contours $V_{i-1}$ and $V_{i+1}$.
    endfor
endwhile
```

Figure 3. ***Left*- Postprocessing method applied k times. At the top, there are snakes produced by Equation (3). The snakes at the bottom are the final results of our system.*Right*-Algorithm for this iteration process**

we postprocess the contours extracted using Equation (3). Let V_i be the optimal snake satisfying Equation (3) for image I_i. Then, we use the snakes V_{i-1} and V_{i+1} to refine the position of snake V_i on image frame I_i. For each element on snake V_{i-1}, we select a number of candidate points along the line between this element and the corresponding element on snake V_i. Among these points, we select the optimal subset minimizing the snake energy. This process is also applied for snake V_{i+1} producing another contour. The resulting two contours are then used for the next iteration of the postprocessing until the snakes find the same position, which is the position of refined snake V_i. If further iterations can not decrease the snake energies before they find the same position, then the weight of similarity energy term is gradually lowered in Equation (2). Notice that if the weight of similarity energy term is zero, both of the snakes are guaranteed to find the same position as long as they have a common subset of candidate points.

Our usage of multiple snakes was motivated by the dual-snake approach proposed by Gunn and Nixon [4], which uses closed snakes unlike our open-ended snakes. In their method, one snake expands from the inside of the region of interest and one snake contracts from the outside. By comparing the total energy of the snakes, the higher energy snake is pushed towards the other by a driving force. The process continues until both snakes find the same position. The method does not explain how to choose the initial positions for the snakes. Our multiple snakes does not use contraction or expansion forces because the tongue surface is not a closed contour. In addition, for our system, the snakes from the adjacent frames automatically provides initial contours.

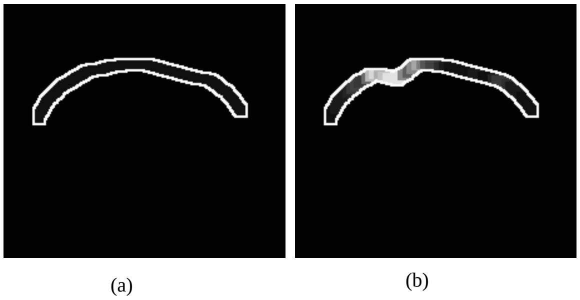
This postprocessing is applied for snakes V_1, V_2, ..., V_{m-1}, which were produced by Equation (3). At the end of this process, we have a new set of the optimal snakes V_1', V_2', V_3', ..., V_{m-1}'. The process continues iteratively until the total energy of optimal snakes does not improve. Figure 3 illustrates the process and shows the basic algorithm. Notice that at iteration 1, refinement of snake V_2' uses contours from V_1, V_2, and V_3. On the other hand, at iteration 2, refinement of snake V_2'' uses contours from V_0, V_1, V_2, V_3, and V_4. As a result, the greater number of iterations means a greater number of contours are included in the estimation of the tongue surface of an image, which was described as one of the requirements for the solution.

3. Estimation of qualitative local deformations

As indicated in Section 2.1, our active contour formulation can be used to measure qualitative local deformations during the movement of the tongue. Although local deformations are not estimated in standard units, knowing what part of the tongue deforms more might be a useful information especially for the diagnosis of Cerebellar Ataxia disorder. Clinical ultrasound examination of tongue movement for Cerebellar Ataxic patients indicates little surface deformation as the tongue moves from one sound to another. Movements are slow and inflexible, with long steady intervals that replace the rapid undulating movements seen in the normal tongue. The posterior and anterior tongue do not interact normally. Posterior tongue posture is abnormally high and rigid resulting in deviant tongue shapes for all sounds. The tip is more normal in behavior. The posterior tongue also moves asynchronously with the anterior tongue. As a result conversational speech is audibly distorted and sounds like alcoholic intoxication. Thus the patient differs from normal in the patterning and timing of the tongue movement.

The simplest method to measure qualitative local deformation is to look at the energy values produced by similarity term $E_{sim}(v_i, S)$ in Equation (2). These values will indicate the similarity of the position of a snake element to the position of the corresponding snake element in model snake or in the previous image's snake. Greater the similarity energy value, more the amount of deformation around that snake element. However this method may have some problems. The position of a snake element is not decided only by the similarity term, but by the combination of internal and external forces. The external force may push the snake element toward a strong edge point, which changes the qualitative local deformation values.

In order to avoid this problem, we calculate the local deformation values by running a separate restricted snake minimization after the actual snake minimization. In the new snake minimization, we do not use the external and the smoothness terms but use only the similarity term. Missing external term means that the restricted minimization will not use any images, which are not necessary because we already know the optimized contour positions. The task of the new restricted minimization is to find the positions of the snake elements on these contours by paying attention only to the similarity energy.


(a)

(b)

Figure 4. Estimation of Qualitative Local Deformations a) The original contour. b) The contour artificially deformed from the contour in (a) by moving one of the snake elements. The gray levels inside the white band represents the amount of local deformation.

Let S be the model snake and let V be the snake extracted by the minimization process for the next image frame. Since V is the optimized snake, we will not change the snake contour of V in the restricted minimization process. In order to find the qualitative local deformation for the i^{th} element of V, we first form the candidate points for i^{th} element by placing equidistant points on the lines between v_{i-1}, v_i, and v_{i+1}. Then, we choose the best similar position by minimizing $E_{sim}(v_i, S)$, where v_i belongs to one of the candidate positions. This process is repeated for all optimal contours using the previous contour in the sequence as a model. Since the candidate points are all on the already detected snake, the newly optimized snake will have the same contour shape as the previous snake. However, the similarity energy values will give us the local deformation amounts. Figure 4-b shows a contour artificially deformed from the contour in Figure 4-a by pulling one of the snake elements. Th contours are depicted as a band for visualization purposes. We run the optimization process for these two contours to detect the local deformation using the contour in Figure 4-a as the model. The gray levels inside the white line in Figure 4-b represents the amount of local deformation; brighter the area, greater the amount of local deformation. Note that the area near the modified snake element has the brightest gray levels and also note that the gray-level variations reflect the deformation for the segments (v_{i-1}, v_i), (v_i, v_{i+1}) from (s_{i-1}, s_i), (s_i, s_{i+1}), respectively.

Our method of extracting local deformations can be compared to matching strategies that defines a number of constraints between the contours segments and uses a minimization to find the optimal match [3]. In our method, the contour patches between the snake elements are analogous to contour segments. The only restriction in the segment matching is the similarity energy unlike Geiger et. al.[3], which has a number of constraints with parameters to fit the requirements of the problem. However, our method produces the qualitative local deformations directly from the optimization which is very similar to the optimization for the contour extraction.

4. Experiments

We tested this system on several different speech and swallowing sequences containing more than 300, 300x300 ultrasound images from HATS. It is essential to note that manual detection of the surface contours for this set takes several days for a speech expert. It took less than one hour for our system to produce the final results on an SGI Octane running one 195 MHz R10000 processor.

Figure 5-a shows the result of a contour extraction-tracking and measurement of local deformation process for a speech sequence. In this sequence, the subject produces the speech sound ”uh gwap”. The user gives the model contour for the first frame, and the rest of the contours are automatically extracted. The qualitative local deformations are marked with gray levels on the same sequence. As indicated in Section 3, the qualitative local deformations are estimated by running a second restricted minimization, which is computationally much less expensive due to reduced number of candidate points and reduced amount of energy calculations. This type of analysis can be helpful to diagnose Cerebellar Ataxia disorder.

During our experiments, we found that our system can handle a wide variety of speech sequences if the initial model is properly chosen. Our current system extends the basic model presented in [1] in order to be able to apply different angular constraints for different tongue movement types. One such movement type is the tongue movement during swallowing. In speech, the tongue surface touches to the top of mount only at a point. For example, in "uh gwap" sequence in Figure 5-a, only the back part of the tongue touches the palate. On the other hand, in swallowing sequences, the whole surface of the tongue may touch the palate, and this causes confusion between the palate and tongue surface. To address these problems, we changed the angle restrictions in the snake minimization. We also chose image frames where the tongue does not touch the palate for the initial model contours. With these modifications, the system was able to detect and track tongue surfaces in swallowing sequences. Figure 5-b shows the analysis of a swallowing sequence, where the subject drinks water. Visual inspection of these contours shows that results are very promising and are virtually identical with the contours extracted manually with the help of previously developed software[9].

5. Conclusions

We presented a system for automatic analysis of 2D contours of the tongue surfaces from digital ultrasound image sequences produced by HATS. This work extends our previous work[1] by applying the system to different speech and swallowing sequences. The enhancements include the addition of angle range constraints for swallowing sequences, qualitative analysis of local deformations to detect special speech disorders, and increased flexibility of the system. We introduce a new active contour model that uses several adjacent images during the extraction of the surface contour for an image frame. We also introduced a discretized-snake internal energy that combines similarity with smoothness. In addition, we used a novel method to decrease the number of candidate points going into the minimization process, which increases the system efficiency.

Our system is very flexible in order to take advantage of speech expert input. The system needs only one initial contour model for the entire image sequence. Although in the literature there are some systems extracting 2D tongue surface contours[7, 9], to the best of our knowledge, our system is the first one to automatically extract surfaces for a complete ultrasound speech sequence. The system can produce qualitative local deformations as a side effect of the minimization process. Our novel internal energy helps us to track the tongue contours, which are open-ended. Open-ended contours are relatively difficult to track because the problem of tracking end points has to be solved. Therefore, systems completely relying on spatiotemporal coherence would fail tracking the end points. Novel combination of smoothness with similarity presented in our paper addresses this problem. Currently, we are working on a new system that uses more spatiotemporal information by fitting a 3D spatiotemporal surface to the entire sequence. This new system can run on a parallel system, taking another advantage of the dynamic programming minimization method.

Acknowledgments

This work is supported by Grant No. R01 DC01758 from NIH and Grant No. IRI 961924 from NSF.

References

[1] Y. S. Akgul, C. Kambhamettu, and M. Stone. Extraction and tracking of the tongue surface from ultrasound image sequences. In *Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition*, Santa Barbara, California, June 1998.

[2] A. A. Amini, T. Weymouth, and R. Jain. Using dynamic programming for solving variational problems in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(9):855–867, 1990.

[3] D. Geiger, A. Gupta, L. A. Costa, and J. Vlontzos. Dynamic programming for detecting, tracking, and matching deformable contours. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17:294–302, 1995.

[4] S. R. Gunn and M. S. Nixon. A robust snake implementation; a dual active contour. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(1):63–68, January 1997.

[5] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, pages 321–331, 1988.

[6] K. F. Lai and R. T. Chin. Deformable contours- modeling and extraction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(11):1084–1090, 1995.

[7] Y. Laprie and M.-O. Berger. Extraction of tongue contours in x-ray images with minimal user interaction. In *Fourth International Conference on Spoken Language Processing*, 1996.

[8] M. Stone and E. Davis. A head and transducer support system for making ultrasound images of tongue/jaw movement. *The Journal of The Acoustical Society of America*, 98(6):3107–3112, December 1995.

[9] M. Unser and M. Stone. Automatic detection of the tongue surface in sequences of ultrasound images. *The Journal of The Acoustical Society of America*, 91(5):3001–3007, May 1992.

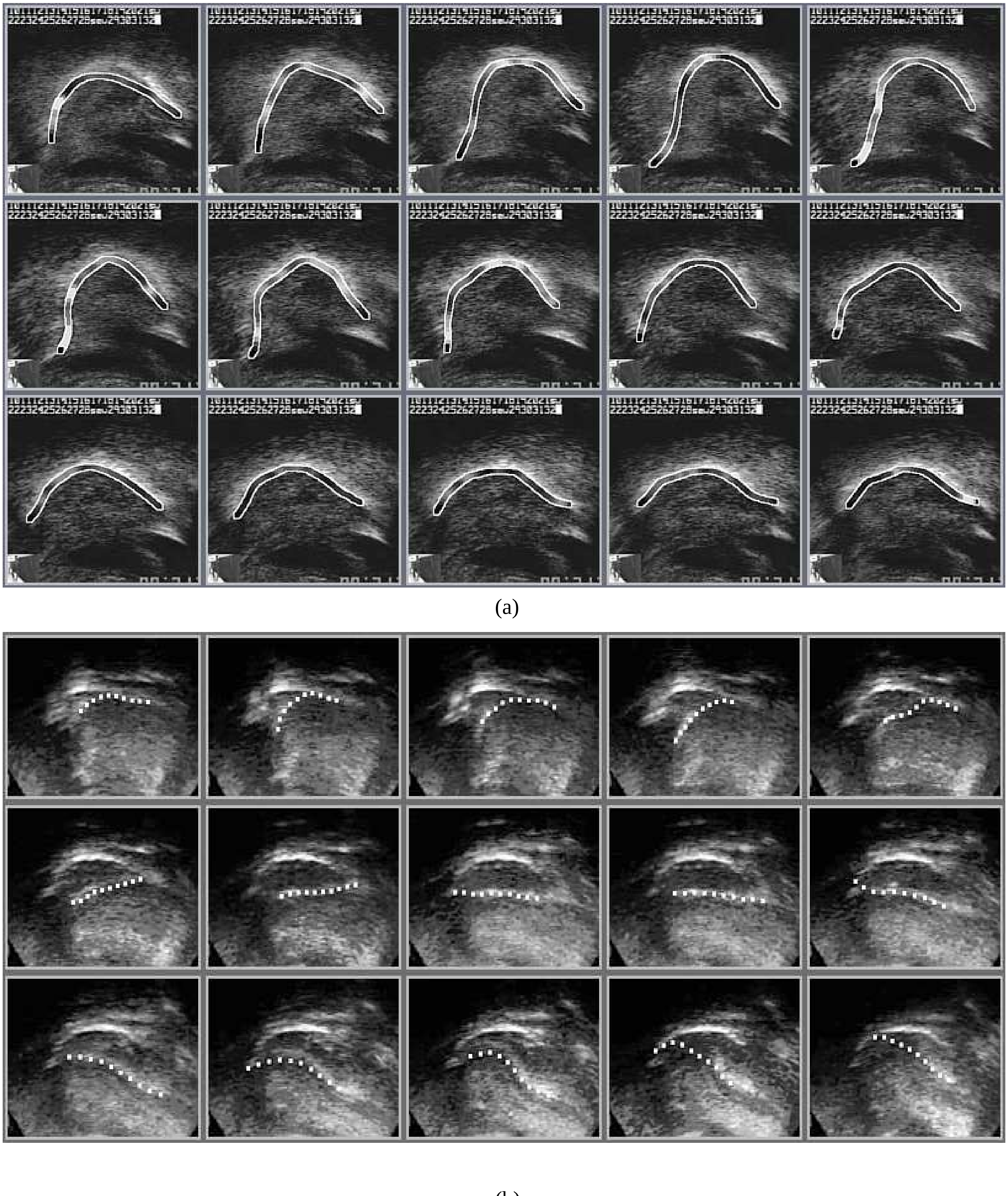


Figure 5. (a) **Detected tongue surfaces of a speech sequence. The subject produces the speech sounds "uh gwap." The tip of the tongue is on the right.** (b) **Analysis of a swallowing sequence, where the subject drinks water. The tip of the tongue is on the left.**