Automatic Extraction and Tracking of The Tongue Contours

Yusuf Sinan Akgul, Chandra Kambhamettu, and Maureen Stone

Abstract— **Computerized analysis of the tongue surface movement can provide valuable information to speech and swallowing research. Ultrasound technology is currently the most attractive modality for the tongue imaging mainly because of its high video frame rate. However, problems with ultrasound imaging such as noise and echo artifacts, refractions, and unrelated reflections pose significant challenges for computer analysis of the tongue images and hence specific methods have to be developed.**

This paper presents a system that is developed for automatic extraction and tracking of the tongue surface movements from ultrasound image sequences. The ultrasound images are supplied by the Head and Transducer Support System (HATS), which was developed in order to fix the head and support the transducer under the chin in a known position without disturbing speech. In this work, we propose a novel scheme for the analysis of the tongue images using deformable contours. We incorporate novel mechanisms to i) impose speech related constraints on the deformations, ii) perform spatiotemporal smoothing using a contour postprocessing stage, iii) utilize optical flow techniques to speedup the search process, and iv) propagate user supplied information to the analysis of all image frames.

We tested the system's performance qualitatively and quantitatively in consultation with speech scientists. Our system produced contours that are within the range of manual measurement variations. The results of our system are extremely encouraging and the system can be used in practical speech and swallowing research in the field of Otolaryngology.

Keywords— **Tongue Motion, Tongue Modeling, Deformable Contours**

I. Introduction

In speech and swallowing research, ultrasound imaging of the tongue is an effective technique to analyze the tongue movement. Analyzing the ultrasound tongue image sequences can provide valuable information for a number of application areas, including disordered speech, aging speech, linguistics, speech processing, and modeling of the tongue. However, manual analysis of the ultrasound tongue images suffers from several drawbacks that is common to many biomedical image analysis areas, which include user-bias, user-fatigue, and not being able to achieve reproducible results[1]. In addition to above common problems, manually processing ultrasound image sequences has other specific problems. The most important problem is the large number of image frames to be analyzed due to the high video capture rates. Analyzing five seconds of speech sequence means extracting contours for 150 images! It is also very difficult for the operator to comprehend the spatiotemporal positions of the tongue contours. As a result, it is crucial to have an automated system for the tongue movement analysis.

Although there are a number of drawbacks, ultrasound technology is the most attractive method of producing a sequence of images of the tongue in motion because it can provide real-time capture rates (30 frames per second), it is non-invasive, convenient for experimentation, and significantly less expensive than other technologies. Alternative methods are too slow to record movement and expensive such as MRI, or they expose subjects to radiation like in X-rays.

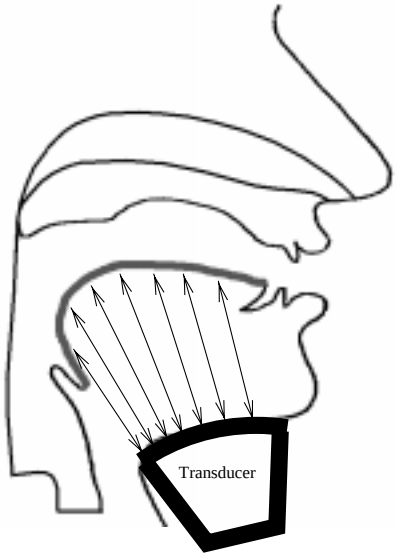


Fig. 1. Position of the ultrasound transducer under chin.

In this paper, we present an extended version of [2] and [3] that extract and track 2D tongue surface contours from ultrasound sequences produced by a Head and Transducer Support System (HATS). HATS[4] was developed for fixing the subject's head while capturing the images, and for supporting the ultrasound transducer under the chin in a known position. These images were tested for reliability and validated. Figure 1 shows the approximate position of the ultrasound transducer with respect to the vocal tract and the tongue. The acoustic waves reflected off the tongue surface will be visualized as depth information proportional to the time delay between the emission and the reception of the ultrasound wave.

Corrupting noise is one of the biggest problems of au-

Yusuf Sinan Akgul and Chandra Kambhamettu are with the Video/Image Modeling and Synthesis (VIMS) Lab, Department of Computer and Information Sciences, University of Delaware, Newark, Delaware 19716. (E-mail: {akgul|chandra}@cis.udel.edu)

Maureen Stone is with the Division of Otolaryngology, The University of Maryland Medical School, Baltimore, Maryland 21201. (E-mail: mstone@umaryland.edu)

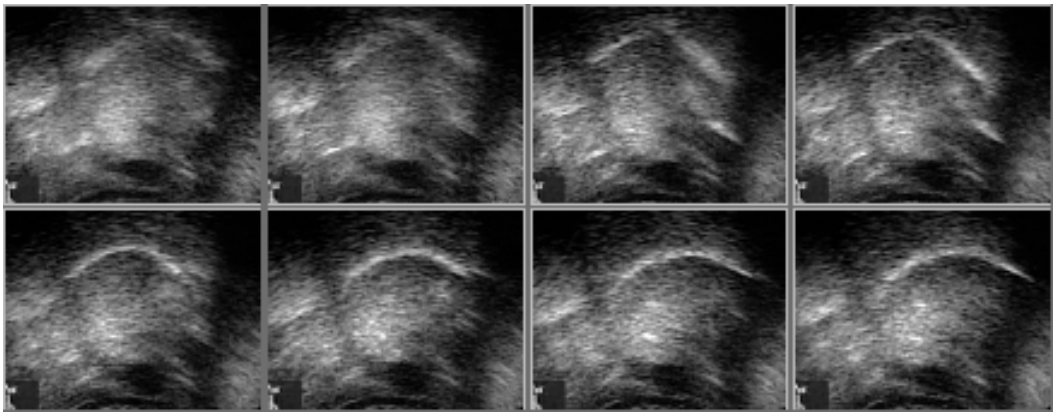


Fig. 3. A subset of a sample ultrasound sequence. The first row shows examples where the tongue contours can only be detected by inspecting the adjacent images in the sequence.

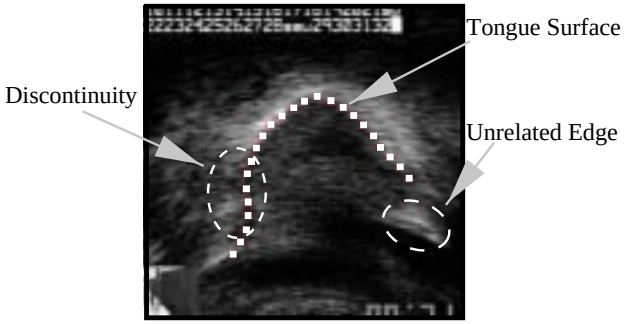


Fig. 2. An example ultrasound tongue image. The tongue surface is automatically detected. The tip of the tongue is to the right. Discontinuities occur when ultrasound waves are close to being parallel to the tongue surface. The waves are reflected but not received by the transducer, causing discontinuities in the final ultrasound image.

tomatic processing of ultrasound images[5]. General ultrasound artifacts[6], such as acoustic shadowing, mirror artifact, and refraction artifact, cause unrelated high intensity areas and discontinuities on the tongue surface. Furthermore, structures within the tongue, such as tendons and blood vessels, may function as good ultrasound reflectors, preventing the passage of the ultrasound wave before reaching the tongue surface. Figure 2 shows an example ultrasound tongue image with the illustrations of some of the problems. Finally, partly as a result of the above problems, in poor images the tongue surface cannot be seen in a single frame, but only in motion. The first row of the Figure 3 shows examples of image frames where the tongue contours can only be detected by inspecting the adjacent images in the sequence.

Our model formulation is based on deformable contours[7], also known as snakes, which are very popular among computer vision and graphics researchers. Many medical applications that require 2D contour or 3D surface analysis use techniques that were popularized by deformable contours. Some of the recent work based on snakes are by Androutsos *et. al.*[8], who use snakes to fit curves for tracer flow extraction in arterial blood flow studies, by Klein *et. al*[9], who employ deformable contours to determine coronary boundaries with discontinuities and branches, and by Yezzi *et. al.*[10], who develop a new snake model that uses image features to drive the optimization process in medical image segmentation.

One of the main reasons of the popularity of deformable contours is their ability to integrate image level bottom up information, task dependent top down knowledge information, and the desirable contour properties into a single optimization process. This is also the main reason we developed our system based on deformable contours. Snakes facilitate initial contours, so that an initial position can be given to the system by the user. We consider this initial contour as an expert knowledge input to the system because it provides the system with the general shape and the location of the tongue surface that the user is searching for. The initial contour is also used to focus only on the related high intensity areas, ignoring the unrelated edges. Interruptions on the tongue contours are common and they pose serious problems in contour extraction. The ability of defining contour shape properties is another attractive feature that we can use to impose smoothness and continuity on the tongue contours. Finally, deformable contours allow additional task dependent constraints that will make the system more efficient and robust.

Among the systems that work on the medical images involving the tongue, to the best of our knowledge, our system is the first to automatically extract and track tongue surfaces for a complete ultrasound sequence. Stone and Lundberg[11] construct 3D tongue surfaces from magnetic resonance images. Unser and Stone[12] use ultrasound tongue images to extract the tongue contours for a single frame. The user has to provide a number of parameters for each frame and there are no explicit models defined for the tongue. Laprie and Berger[13] also use snakes to extract tongue surfaces from X-ray images. Although similarities exist between our systems, there are fundamental differences in terms of imaging modalities and deformable contour models. Their specific problems are very different than

ours. For example, in X-rays the discontinuities are caused by other structures such as teeth and fillings. On the other hand, discontinuities in our system is caused by mainly refractions, and reflections from a non-perpendicular interface, which are relatively more difficult to handle.

The remainder of this paper is organized as follows. The next section discusses our formulation of the deformable model. Section III discusses how we used our deformable model in tracking the tongue contours. Section IV develops an energy minimization scheme to optimize the deformable contour while satisfying the spatiotemporal constraints. Section V introduces the methods developed for incorporating domain dependent knowledge into the contour extraction and tracking process. Section VI describes the experiments that we performed to validate our system's performance.

II. The Deformable Model

We define our deformable model by a discrete version instead of the original continuous formulation. The motivation of using a discrete formulation will be explained at the end of this section. A deformable contour model is a set of ordered discrete points $V = [v_1, v_2, ..., v_n]$ with an energy functional that is minimized on an image frame I with a given initial model contour $S = [s_1, s_2, ..., s_n]$.

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

where α and β are the weighting parameters, $E_{int}(v_i, S)$ is the internal energy and $E_{ext}(v_i, I)$ is the external energy.

A. Internal Energy

$E_{int}(v_i, S)$ is the internal energy of v_i which is the weighted sum of smoothness $E_{smo}(v_i)$, similarity to the initial model $E_{sim}(v_i, S)$, and a distance energy $E_{dist}(v_i)$, which keeps the snake elements at equidistant intervals.

$$E_{int}(v_i, S) = \overrightarrow{\lambda} \cdot \begin{bmatrix} E_{smo}(v_i) \\ E_{sim}(v_i, S) \\ E_{dist}(v_i) \end{bmatrix}. \qquad (2)$$

$\overrightarrow{\lambda} = [\lambda_1\ \lambda_2\ \lambda_3]$ is a vector containing weighting parameters for the internal energy terms, where $0 \le \lambda_1, \lambda_2, \lambda_3 \le 1$ and $\lambda_1 + \lambda_2 + \lambda_3 = 1$.

$E_{smo}(v_i)$ is similar to the smoothness term in the original snake formulation[7]. However, it measures only the bending amount as opposed to the second derivative of the contour. It is estimated as (Figure 4),

$$E_{smo}(v_i) = 1 - \cos\theta_i = 1 - \frac{\overrightarrow{v_{i-1}v_i} \cdot \overrightarrow{v_iv_{i+1}}}{|\overrightarrow{v_{i-1}v_i}|\,|\overrightarrow{v_iv_{i+1}}|}$$

where $i = 2 \ldots (n-1)$. We define $E_{smo}(v_1) = E_{smo}(v_2)$ and $E_{smo}(v_n) = E_{smo}(v_{n-1})$. Figure 5 shows how the energy produced by the smoothness term changes with respect to the angle θ between the vectors $\overrightarrow{v_{i-1}v_i}$ and $\overrightarrow{v_iv_{i+1}}$. As apparent from the figure, $E_{smo}(v_i)$ tolerates small bendings. However, it increases steeply if the angle θ increases.

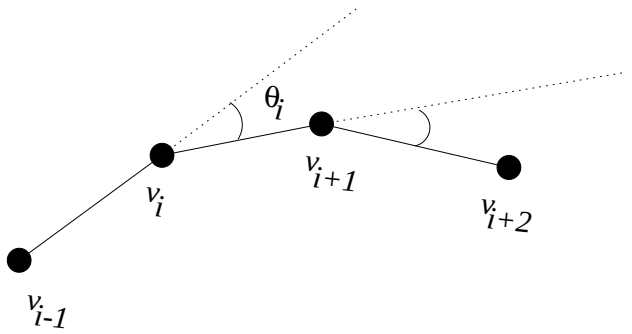


Fig. 4. Smoothness term: Amount of bending of the contour at v_i.

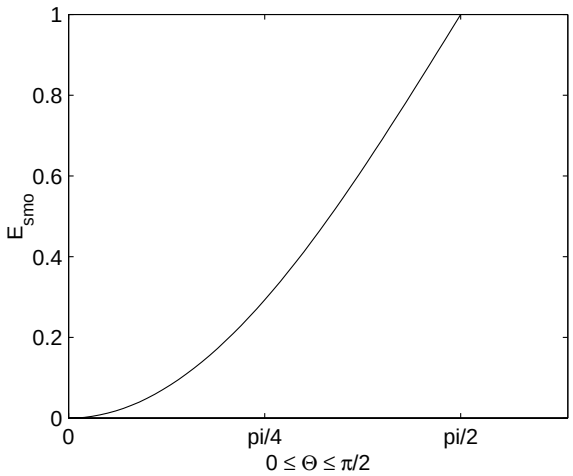


Fig. 5. The change of $E_{smo}(v_i)$ with respect to θ.

The similarity term $E_{sim}(v_i, S)$ has a crucial role in our model. The main function of similarity term is to propagate contour information provided by the speech scientist to the surface extraction processes of all image frames. This task is achieved by forcing the newly extracted contours to be similar to the user-supplied-contour by minimizing the bending changes between contours. Given a model contour S, the similarity term $E_{sim}(v_i, S)$ is given by Equation(3).

If the newly extracted contour tries to increase the bending amount relative to the model S, then both the similarity and the smoothness terms try to decrease the bending amount. However, if the newly extracted contour tries to decrease the bending amount, then the similarity and the smoothness terms are in conflict because the smoothness term forces the contour to be a straight line, whereas the similarity term forces the contour to increase bending to have the exact shape of the given model S.

The distance energy $E_{dist}(v_i)$ keeps the snake elements at equidistant intervals,

$$E_{dist}(v_i) = \left| \frac{|v_i - v_{i-1}|}{\frac{1}{n-1}\sum_{j=2}^{n} |v_j - v_{j-1}|} - 1 \right|.$$

Intuitively, the energy term of original snake formulation[7] that imposes flexibility on the contour can be considered as the combination of our smoothness and distance energies.

$$E_{sim}(v_i,S)=\left|\left(1-\frac{\overrightarrow{v_{i-1}v_i}\cdot\overrightarrow{v_iv_{i+1}}}{|\overrightarrow{v_{i-1}v_i}|\,|\overrightarrow{v_iv_{i+1}}|}\right)-\left(1-\frac{\overrightarrow{s_{i-1}s_i}\cdot\overrightarrow{s_is_{i+1}}}{|\overrightarrow{s_{i-1}s_i}|\,|\overrightarrow{s_is_{i+1}}|}\right)\right|=\left|\frac{\overrightarrow{s_{i-1}s_i}\cdot\overrightarrow{s_is_{i+1}}}{|\overrightarrow{s_{i-1}s_i}|\,|\overrightarrow{s_is_{i+1}}|}-\frac{\overrightarrow{v_{i-1}v_i}\cdot\overrightarrow{v_iv_{i+1}}}{|\overrightarrow{v_{i-1}v_i}|\,|\overrightarrow{v_iv_{i+1}}|}\right|. \quad (3)$$

By decomposing this energy into two parts, we have a better control on the snake properties. For example, we may choose to relax the bending constraints without affecting the distance energy. We take advantage of this formulation in incorporating the task dependent knowledge into the optimization process, which is explained in Section V.

One may argue that the increased number of functions in our internal energy term introduces new weighting parameters, λ_1, λ_2, and λ_3, thus our formulation is more sensitive to selection of the weighting parameters. However, it is not the case. In Equation (2), the weighting parameters are automatically set using the *min-max* principle introduced by Gennert and Yullie[14] for determining optimal weights in multiple function optimization. Although we still need to choose a value for the parameter α and normalize the internal energy terms, employment of the min-max principle makes our system robust against parameter selection. The advantage of using the min-max principle is that it will prevent the selection of a weighting parameter vector $\overrightarrow{\lambda}$ that represents $E_{smo}(v_i)$, $E_{sim}(v_i,S)$, or $E_{dist}(v_i)$ insufficiently in the minimized internal energy. Min-max principle achieves this by maximizing the Equation (2) with respect to $\overrightarrow{\lambda}$ and minimizing it over possible v_i positions. It can be shown that, for a given v_i position, maximizing over $\overrightarrow{\lambda}$ in Equation (2) is equivalent to choosing the maximum of $E_{smo}(v_i)$, $E_{sim}(v_i,S)$, and $E_{dist}(v_i)$[15] [14]. Intuitively, this makes sense, because if one of these terms is getting high for an image region, it would be desired that this high value is represented more effectively in the overall minimization process to enforce more smoothness or similarity or distance force, depending on the maximum valued term.

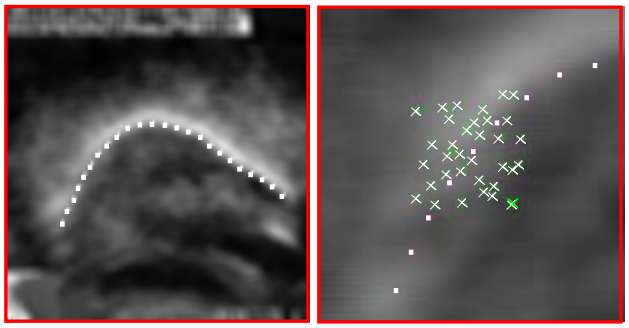
Although we found the min-max principle useful in Equation (2), it did not perform well for the weight parameters of Equation (1). This was because, in ultrasound tongue images, image intensity of some portions of the tongue surface is interrupted. For these image regions, the external energy term of Equation (1) will always be high because the image gradient for these areas is almost zero. Therefore, the min-max algorithm will use this higher valued external energy term over the internal energy term in the minimization process, which is the opposite action of that needed for regions of this type. Internal energy should be the deciding force in this situation. As a result, the min-max algorithm fails to choose reasonable weight parameters for the minimization process in Equation (1). This is the reason we used different weight parameter mechanisms for Equation (1) and Equation (2).

B. External Energy

The external energy term $E_{ext}(v_i,I)$ is given by the negative of the image gradient at v_i. Although the formulation in Equation(4) for the external energy may not look robust for our specific application, in this section we explain how we implement a robust external energy functional by considering the tangent angle of the contour and image gradient value.

$$E_{ext}(v_i,I)=-\left|\nabla I(v_i)\right|. \quad (4)$$

Our method uses a coarse-to-fine strategy. At the coarsest level, the system works on a useful subset of available pixels. The selection criteria of this subset is chosen considering the effects on the external energy because we know that only the external energy will use the elements of this subset. Let V_0 be the initial user supplied contour on image frame I_0. Let (x_0,y_0) be the position of the first element v_0 of V_0 and let W_0 be the area of interest for v_0 on image I_1. Then we can choose the most useful elements of W_0 in terms of better optimization by looking at the similarity of the angles of gradient values $\nabla I_0(x_0,y_0)$ and $\nabla I_1(x_1,y_1)$, where (x_1,y_1) is an element of area of interest W_0. This method uses the basic principle of optical flow techniques which assume that intensity of a given point does not change with respect to time[16]. Using this principle, it is assumed that the above values $\nabla I_0(x_0,y_0)$ and $\nabla I_1(x_1,y_1)$ are close. One improvement on this method is to use the normal direction of the tongue contour at the snake element v_0 of V_0 instead of $\nabla I_0(x_0,y_0)$. This is a better choice because we know that the ideal value of the $\nabla I_0(x_0,y_0)$ should be the normal direction of the tongue surface contour at position (x_0,y_0). Therefore, using the normal direction instead of gradient value produces better results.



(a) (b)

Fig. 6. (a) Snake V_0 in image I_0. (b) Scaled up view of selected candidate points (crossed) for one of the elements of V_0 on image I_1.

Our experience shows that selecting l to $2l$ elements from a l by l search-window using the above method can produce acceptable results, in addition to the reduced computational complexity.

However, this method has one disadvantage: the basic principle of optical flow may not hold for some instances of our problem. In ultrasound tongue images, some areas of the tongue surface may not have any intensity contrast as shown in Figure 2. If the element v_0 is in such a region, then selecting the points that has similar gradient angles in area of interest W_0 will not produce any useful points for snake minimization because the intensity value at the position of v_0 on image I_0 is not relevant to the tongue surface. To address this problem, we divided the area of interest W_0 into l different disjoint areas and we selected the most useful point from each disjoint area as our best element of W_0. This way, even if the similar points do not have useful gradient values for snake minimization, we expect the internal forces to drive minimization process to impose desired contour properties. Figure (6-b) shows the selected points for a portion of the snake in Figure (6-a).

As we mentioned in Section I, the power of deformable contour framework comes from its ability to integrate domain dependent knowledge with the desirable contour attributes and the image features. In order to integrate high level knowledge into the optimization process, researchers almost always modify the snake mechanism to fit the needs of their specific application. In fact, it is very rare that two different snake based systems use the same snake formulations even if they operate on the same task. For example, although the systems of Berger *et. al.*[17] and Chalana *et. al.*[18] both extract left ventricle boundaries from echographic images, they prefer to use very different snake mechanisms. Berger *et. al.* modify the snake mechanism so that the external energy is utilized in a way that tracking of ventricle is improved. On the other hand, Chalana *et. al.* modify the snake mechanism so that more spatiotemporal information is used in the snake minimization process. Both modifications are due to the researchers motivations and the problems addressed.

For the same reason, our snake formulation for the deformable contours differs from the other formulations. One of the main differences of our formulation is that it uses a similarity term to properly utilize user supplied contour information in the tracking process. This term is also used as a way of integrating spatiotemporal information into the extraction of a single tongue contour. Our system needs this internal energy term because, as described in Section I, for some poor quality ultrasound tongue images, the system needs to rely on adjacent images or the user supplied contour to be able to reliably extract the tongue contour. With the similarity term, we extend the deformable contours so that they fit the needs of our problems.

Another difference of our snake formulation is its discrete internal energy term and the methods we use to minimize the snake. A discrete internal energy term provides us more controllability of the internal energies and a way of automatic weighting parameterization. Although, these types of features can be obtained using a continuous internal energy, such as B-Snakes as in Klein *et. al.*[9] and Blake and Isard[19], they would interfere with our coarse to fine optimization methods as described in Section II-B. Our external energy term depends on the individual pixels that are selected using optical flow techniques. Since a B-Snake does not have to interpolate a given control point, we decided to develop our internal energy formulation as discrete terms. Discrete energy terms are used by many snake based systems; one recent example is Malassiotis and Gerassimos[20], who use a discretized snake energy formulations for tracking the left ventricle enhanced by a temporal learning-filtering process. As explained in Section IV, there are very effective methods for the optimization of these types of deformable contours.

III. Tracking the Tongue Contours

Tracking of the tongue contours are handled by using the usual deformable contour tracking method first proposed by Kass *et. al.*[7] and Terzopoulos *et. al.*[21]. In this approach, the model provided by the user is used as the initial position for a contour optimization problem. The optimized contour is used as the initial position for the optimization of the next frame, and the process continues until all the images are processed. We express the above scheme formally as follows.

Let $V = [v_1, v_2, ..., v_n]$ be a snake, and W_i be an l by l *search-window* centered at the position of v_i on image frame I for $i = 1..n$. Let $Y = [y_1, y_2, ..., y_n]$ be another snake where $y_i \in W_i$ for $i = 1..n$. Then, we define *the set of possible snakes*, $P(V, I)$, as the set of all possible Y's (Figure 7).

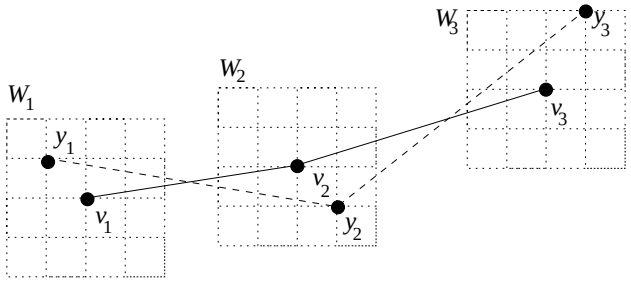


Fig. 7. $Y' = [y_1, y_2, y_3]$ is an element of $P(V', I)$, where $V' = [v_1, v_2, v_3]$ and W_i is a 5 by 5 search-window. The number of elements in $P(V', I)$ is $l^{2^n} = 5^{2^3}$.

Using the above definition of $P(V, I)$, we write the total energy of an optimized set of snakes for an ultrasound image sequence $I_0, I_1, ..., I_m$ as:

$$E_{Sequence} = \min \sum_{k=1}^{m} E_{Snake}(V_k, V_{k-1}, I_k) \qquad (5)$$

where $V_k \in P(V_{k-1}, I_k)$ and V_0 is the initial contour given by the user for image frame I_0.

Notice that the optimization on an image frame I_k is linked to the optimizations on frames I_{k-1} and I_{k+1} by contours that are in a set of possible snakes. The final solution to our problem is a set of snakes that satisfy Equation (5). Intuitively, this optimization process of a set of snakes can be represented by a tree (Figure 8). The initial contour, V_0, given by the user is placed at the root of the tree.

The members of the set of possible snakes produced by V_0, $P(V_0, I_0)$, become the children of the root. Then we form possible sets of snakes for each child node of the root and the resulting members become the children of that node. The process continues similarly for m levels where m is the number of image frames in a given sequence. The final solution to extraction and tracking of the tongue surfaces for a sequence can be established by finding a path from the root node to one of the nodes at the lowest level satisfying the Equation (5). Exhaustively checking each possible path is computationally very expensive since there are exponential number of such paths. The next section explains how we approximate the Equation (5) with our implementation details.

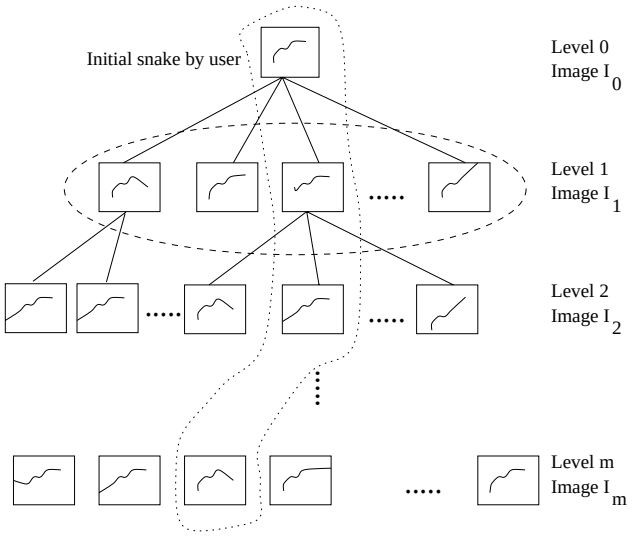


Fig. 8. A tree can represent the estimation of the contours for a sequence of images. (The nodes surrounded by the dotted line are the optimal snakes. The dashed line surrounds the elements of $P(V_0, I_0)$.)

IV. Energy Optimization

Finding a global minima for the energy functional of Equation(5) is a difficult task. We already mentioned that iterating on the tree paths of Figure 8 is computationally prohibitive. Common snake minimization schemes such as the one by the original proposal[7] or by Cohen[22] are based on gradient descent search and they are sensitive to local minima, which is a serious problem in ultrasound images due to the high amount of noise. Open contours are another problem with these methods because their basic formulations assume closed contours. Furthermore, these methods do not easily allow hard constraints, which are the type of rules in the optimization process that can never be violated.

There is another group of deformable contour optimization techniques that are based on application of dynamic programming to the discretized energy functionals, such as ours. Amini *et. al.*[23] introduced a dynamic programming based minimization method that finds optimal results in polynomial time. Geiger *et. al.*[24] proposed another dynamic programming based method that uses gradient continuity to impose smoothness. Although it does not use dynamic programming techniques, Williams and Shah[25] proposed a non-optimal greedy algorithm for snake minimization that is similar to the first two methods. It is trivial to impose hard constraints with the above three methods. In addition, the above methods are numerically stabler than the gradient descent based methods.

Our system is based on the method by Amini *et. al.*[23]. We approximate the optimization in Equation (5) by

$$E_{Sequence} = \sum_{k=1}^{m} \min E_{Snake}(V_k, V_{k-1}, I_k) \qquad (6)$$

where $V_k \in P(V_{k-1}, I_k)$ and V_0 is the initial contour given by the user.

In terms of tree operations, Equation (6) implies that only the children of the minimum energy snake are produced for the lower tree level. Although this modification reduces the computational cost to a feasible level, it does not necessarily produce an optimal solution, which is guaranteed by Equation(5). Intuitively, the problem is that while minimizing the snake for tree level i, the method does not get any image or contour information from tree level $i+1$ and deeper tree levels. This results in lack of spatiotemporal consistency of the optimized contours, which may create serious problems for our system because during the contour extraction process, the system should use as much information as possible from the adjacent frames in order to locate the tongue surface better. We should note that checking the adjacent images is compulsory even for human speech scientists for a successful analysis. The fourth row of Figure 13 shows examples of image frames where the tongue contours can only be detected by inspecting the adjacent images in the sequence.

A. A contour postprocessing technique to impose spatiotemporal information

In order to overcome the problem of lack of spatiotemporal consistency on the contours, we postprocess the contours extracted by Equation (6). The postprocessing is done by a separate snake optimization that includes another internal energy term, E_{stc}, to impose spatiotemporal constraints by forcing smoothness of the interframe motion of each snake element.

$$E_{stc}(v_i^k) = 1 - \frac{\overrightarrow{v_i^{k-1}v_i^k} \cdot \overrightarrow{v_i^k v_i^{k+1}}}{|\overrightarrow{v_i^{k-1}v_i^k}|\,|\overrightarrow{v_i^k v_i^{k+1}}|}$$

where $k = 1 \ldots (m-1)$, m is the number of frames in a sequence, and v_i^k is the i^{th} element of the contour for image frame k.

Let V_k be the optimal snake satisfying Equation (6) for image I_k. Then, we use the snakes V_{k-1} and V_{k+1} to refine the position of snake V_k on image frame I_k. For each element on snake V_{k-1}, we select a number of candidate points along the line between this element and the corresponding element on snake V_k. Among these points, we

select the optimal subset minimizing the snake energy with the E_{stc} term. This process is also applied for snake V_{k+1}. The process continues iteratively until the snakes find the same position, which is the position of the refined snake V_k. If further iterations cannot decrease the snake energies before they find the same position, then the similarity energy term is gradually relaxed in Equation (1). Notice that if there are no similarity energy terms, both snakes are guaranteed to find the same position as long as they have a common subset of candidate points. Figure 9 illustrates this process.

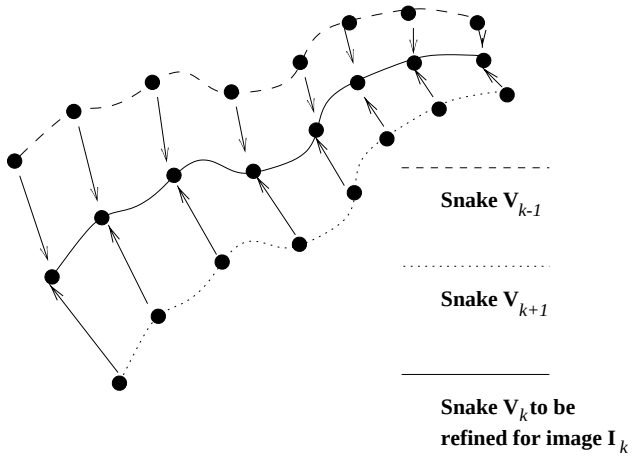


Fig. 9. Refining the position of snake V_k.

The postprocessing is applied for snakes V_1, V_2, ..., V_{m-1} produced by Equation (6). 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 the optimal snakes do not improve. Figure 10 illustrates the process. 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 should impose spatiotemporal constraints to the general optimization process.

Our utilization of multiple snakes resembles the dual-snake approach proposed by Gunn and Nixon [26], who use closed snakes on static images unlike our open-ended snakes in image sequences. 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 do 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.

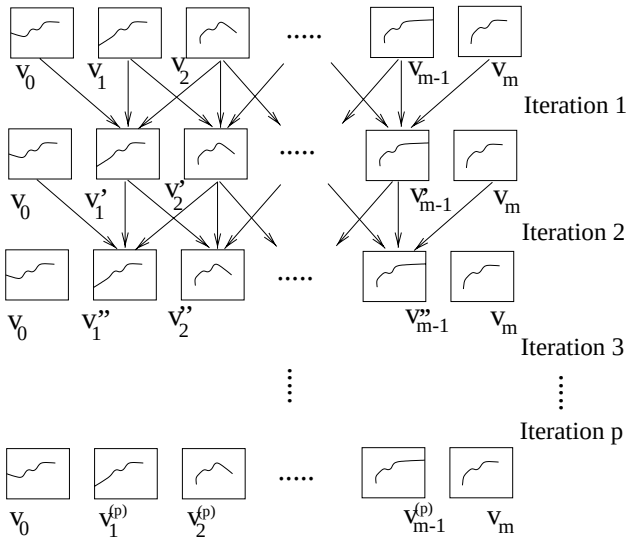


Fig. 10. Postprocessing is applied p times. At the top, there are snakes produced by Equation (6). The snakes at the bottom are the final results of our system.

V. Incorporating domain dependent knowledge into the optimization process

In medical image analysis, general information about the object in study is usually available by means of medical literature or medical experts. We know that although tongue is a completely nonrigid structure, there are some impossible positions that it cannot take, which can be built into the deformable model as constraints to increase the robustness of the automatic analysis. For example, the tongue surface can reflect ultrasound waves only in a range of angles. We use this information in our external energy to ignore some gradient values in choosing the most useful pixels as explained in Section II-B. Another domain dependent constraint is to use angular velocity of the tongue surface tangents. Although the surface velocity of the tongue is very high even for a 30 frame per second capture rate, we can impose constraints on the angular velocity of the surface tangent. That is, we can impose a constraint on the maximum change in the direction of the vector $\overrightarrow{v_{i-1}v_i}$ between two frames. We found out that this change is not more than 35 degrees for the most speech applications. These general domain dependent constraints are incorporated in the snake optimization process as hard constraints. The discretized nature of our optimization process makes it very convenient to reject the tongue positions that violate the constraint rules. Our formulation of the main snake energy was specifically designed to take advantage of feasibility of hard constraints on discretized energy formulations.

In addition to the above general constraints, we can impose other constraints depending on the speech experiment type. *Phonetics* is the study of speech sounds by classifying the spoken sounds according to the way they are produced by the organs of speech[27]. Although many organs are involved in sound production process, such as lips, teeth, palate, nasal cavity, etc., the tongue has the most impor-

TABLE I
TONGUE POSITIONS DURING THE PRODUCTION OF SOME ENGLISH VOWELS. THE VOWELS ARE THE CAPITALIZED PARTS OF THE WORDS.

Vowels

	Tongue Part		
	Front	Central	Back
High	"sh**EE**p" "sh**I**p"		"b**OO**t" "p**U**t"
Middle	"b**E**d"	"cupb**OAR**d" "b**IR**d"	"p**O**t" "c**AUGH**t"
Low	"b**A**d"	"c**U**t"	"c**AL**m"

tant function. As a result, many classification schemes are based on the positions of the tongue parts. Figure 11 shows the position of the tongue inside the vocal tract during the production of three English vowels.

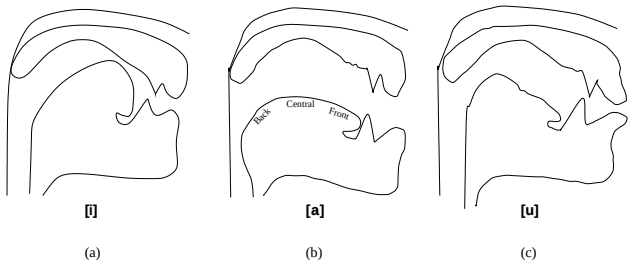


Fig. 11. Position of the tongue inside the vocal tract while producing English vowels (a) "sh**EE**p" (b) "c**AL**m" (c) "b**OO**t".

In Phonetics, the vowels are generally classified by the positions of the front, the middle and the back part of the tongue, because shaping of the vowel is mostly achieved by the tongue. Table I shows classification of some of the English vowels.

Diphthongs are single phonemes that have a sequence of two different vowel positions[28]. One example diphthong occurs in "c**A**ke", where the front part of the tongue moves from middle position to high.

Often, speech research requires experiments on production of vowels, diphthongs, or transitions between them. Such experiments can provide valuable information for a number of studies, such as the effects of aging in speech production and detection of tongue related speech disorders. Since we know what the subjects say during these experiments, we expect the tongue to take anticipated shapes by using the phonetic classification of the utterances. We take advantage of this additional domain dependent information by constraining the possible deformable model shapes by utilizing hard constraints in our optimization process. We specify the constraints by forcing the parts of the tongue (back, middle, and front) to take surface tangent angles in a specified range depending on the produced sound. Figure 12 shows two automatically tracked speech sequences with different tongue motion types and different surface angle constraints. The sequence in Figure (12-a) is a front raising speech. The sequence in Figure (12-b) is a middle raising speech.

Using domain dependent information in the optimization process makes the system both efficient and robust. However, if we do not have information that can constrain the tongue surface shapes, the system still can track the tongue contours by increasing the number of candidate points in the search window during the optimization. In other words, the system can trade efficiency for robustness if there is not enough constraints on the tongue shapes. Figure 13 shows a speech sequence with an unknown speech type. Figure 14 shows the tongue surface contours for the same sequence, which are automatically extracted without using any speech type constraints.

VI. EXPERIMENTS AND VALIDATION

We performed extensive testing experiments on a number of image sequences. The system can extract the contour of a frame in less than 10 seconds on an SGI Octane running one 195 MHz R10000 processor. It is essential to note that manual detection of the surface of a single frame takes several minutes for a speech scientist. Needless to say, manually analyzing long sequences of tongue images, which is the common case, decreases both the performance and efficiency.

We first validated our results with visual inspection and qualitative analysis by the speech scientists, which showed that the results are precise and reliable for speech and swallowing research. For the visual inspection and qualitative analysis validation, we used about 20 sequences. Although lack of ground truth makes the numerical comparisons very difficult, we performed experiments that numerically compare the automatically extracted contours with the manually detected contours. Two speech scientists independently extracted the contours for two 15-frame speech sequences manually. One sequence is a typical case where tongue surface discontinuities are minimal (such as the last row of Figure 13). The other sequence has discontinuities at one end of the tongue surfaces, thus is it relatively difficult to analyze this sequence. We then measured a mean sum of squared distances(MSSD) between the two manually detected sequences, and between the automatic sequence and the manually detected sequences; we have normalized MSSD values for visualization purposes(NMSSD).

During the calculation of the MSSD values, we measured the distances between closest snake elements of each contour. Formally, given two deformable contours $V = [v_1, v_2, ..., v_n]$ and $U = [u_1, u_2, ..., u_n]$, the MSSD value between these contours is calculated as

$$MSSD(U,V) = \frac{1}{2n}\left(\sum_{i=1}^{n}\min_j (v_i - u_j)^2 + \sum_{i=1}^{n}\min_j (u_i - v_j)^2\right)$$

which is similar to what Chalana and Kim[29] suggested for evaluation of boundary detection on medical images. This measure is also similar to what Unser and Stone[12] use to validate their extracted tongue contours.

Figure 15 shows the NMSSD values for the second sequence, which is a difficult test case. As seen on this figure, the NMSSD values are relatively higher at the end of

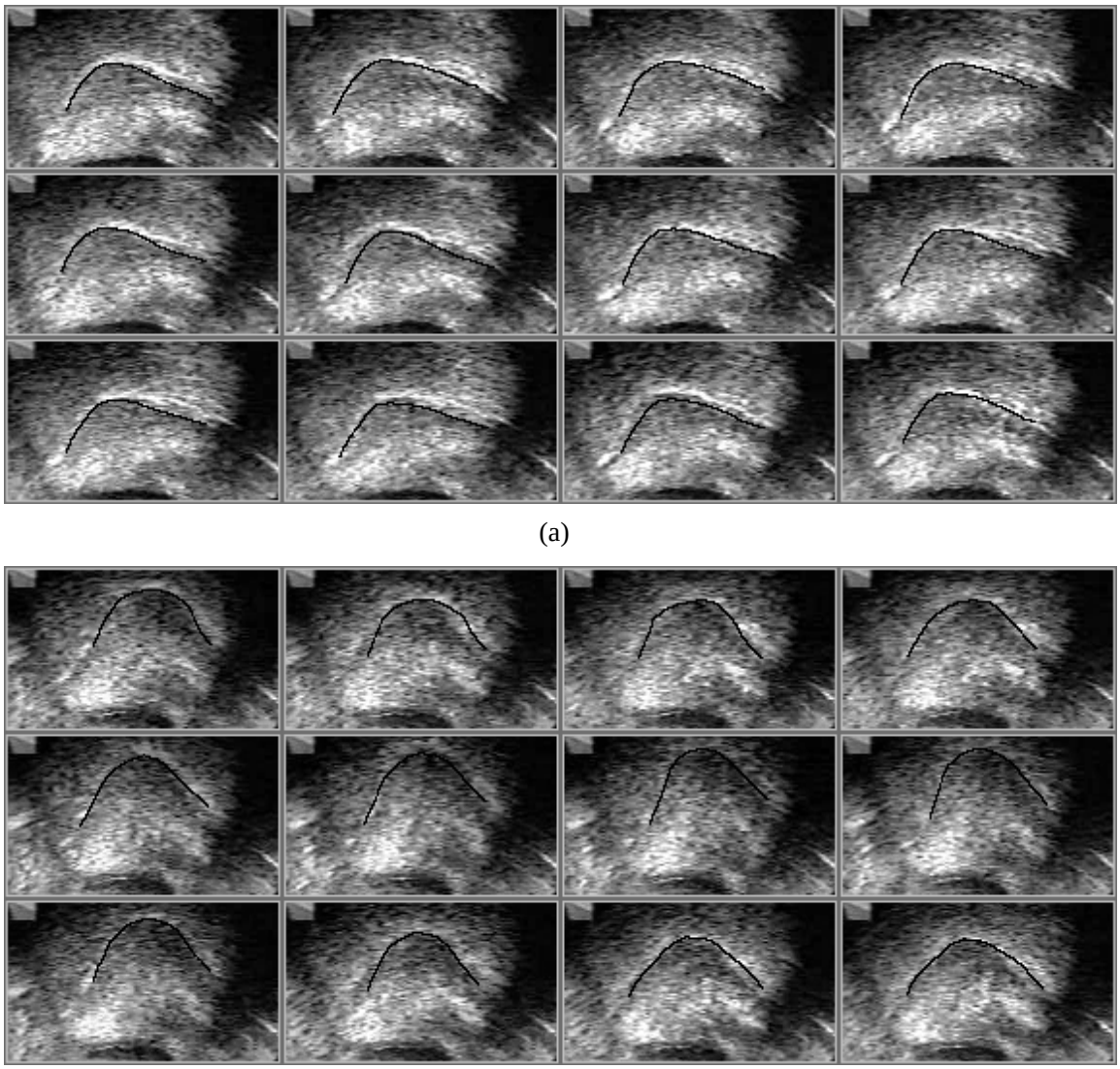


Fig. 12. (a) A front raising sequence "dadada" (b) A middle raising sequence "gagaga". (The tongue tips are to the right.)

the sequence, where the surface discontinuity becomes a serious problem and the scientists had to use their knowledge and past experiences to estimate the tongue position at the discontinuities. Figure 16 shows the NMSSD values for the typical sequence, where NMSSD values are within a reasonable range. Table II shows the MSSD values for the two test cases. It can be observed that the MSSD values are consistent in the typical test case. On the other hand, for the difficult test case, the MSSD values in general are higher. This is because even the speech scientists had difficulty in concurring to similar image positions that describe the tongue contour, due to the lack of tongue surface information in problematic areas.

The results of our qualitative and quantitative experiments showed that the automatically detected contours are within the range of manual measurement variation between the scientists, justifying the practicality of our system in tongue research. The system of Unser and Stone[12], which operates only on individual frames and is semi-automatic, reports a similar performance.

VII. Conclusions

We presented a system for the automatic analysis of the tongue surfaces from digital ultrasound image sequences produced by HATS. HATS was developed for reliable and valid ultrasound imaging of the tongue. Our system is the first to automatically extract and track the tongue contours from an ultrasound sequence. The presented system is valuable for speech and swallowing research.

Our system uses a discrete formulation for the deformable contour model as the main analysis technique. We introduced a number of novel ideas that can be useful in other areas of medical image analysis. First, in order to make system more robust and efficient, we imposed speech, tongue and ultrasound imaging constraints in our analysis. Second, we introduced a new contour postprocessing tech-

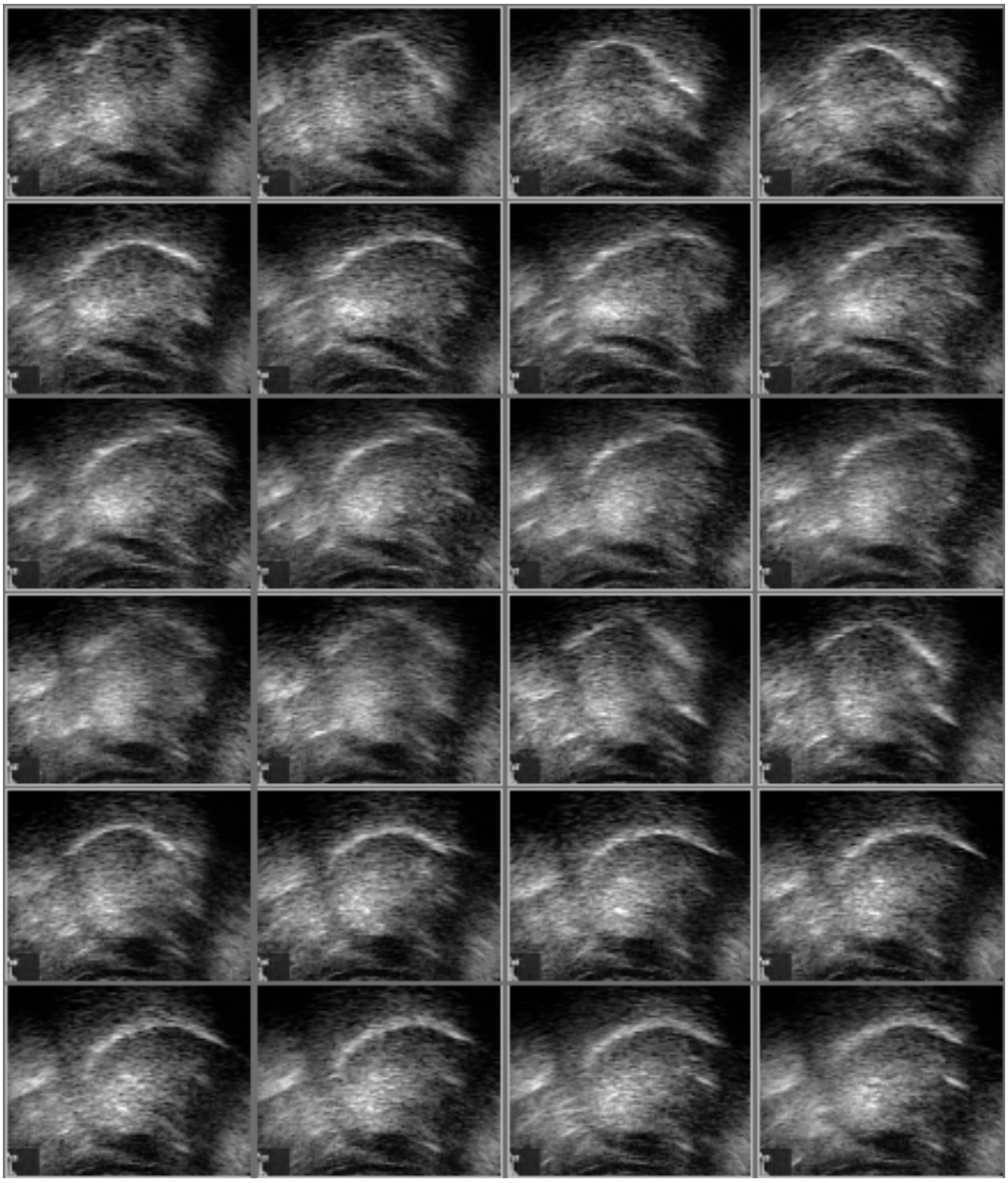


Fig. 13. A speech sequence with an unknown speech type.

nique in order to make the tongue surfaces spatiotemporally consistent. Third, we extended the deformable contour model in a way that it uses the user supplied initial contour as an expert knowledge input in order to help the system to detect tongue contours where there is not enough image data.

We tested the system's performance visually and numerically in consultation with speech scientists. Our system produced contours that are within the range of manual measurement variations. Given the nature of the problem, the results of our system are extremely encouraging and it can be used in practical speech and swallowing research.

Acknowledgments

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

References

[1] T. McInerney and D. Terzopoulos, "Deformable models in medical image analysis: A survey," *Medical Image Analysis*, vol. 1, pp. 91–108, 1996.

[2] Yusuf Sinan Akgul, Chandra Kambhamettu, and Maureen Stone, "Extraction and tracking of the tongue surface from ultrasound image sequences," in *Proceedings of The IEEE Computer Vision Pattern Recognition*, Santa Barbara, California, June 1998, pp. 298–303.

[3] Yusuf Sinan Akgul, Chandra Kambhamettu, and Maureen Stone, "Automatic motion analysis of the tongue surface from ultrasound image sequences," in *Proceedings of The IEEE Workshop on Biomedical Image Analysis*, Santa Barbara, California, June 1998, pp. 126–132.

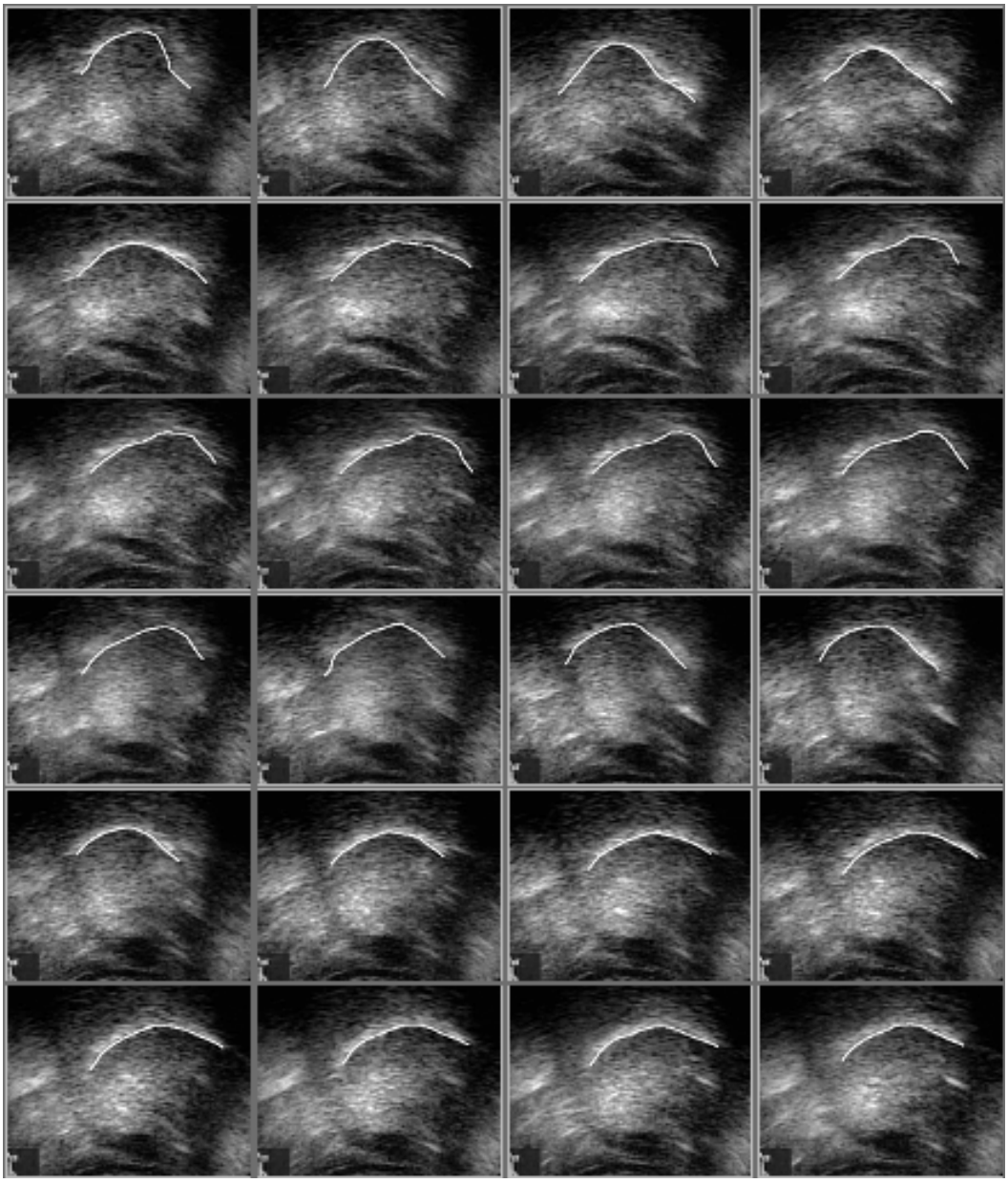


Fig. 14. Automatically tracked contours for Figure 13.

[4] M. Stone and E.P. Davis, "A head and transducer support system for making ultrasound images of tongue/jaw movement," *The Journal of The Acoustical Society of America*, vol. 98, no. 6, pp. 3107–3112, December 1995.

[5] N. Ayache, I. Cohen, and I. Herlin, "Medical image tracking," in *Active Vision*, A. Blake and A. Yuille, Eds., pp. 285–301. MIT Press, 1992.

[6] D. Pickuth, *Essentials of Ultrasound*, Springer-Verlag, 1995.

[7] M. Kass, A.P. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International Journal of Computer Vision*, vol. 1, no. 4, pp. 321–331, January 1988.

[8] D. Androutsos, P.E. Trahanias, and A.N. Venetsanopoulos, "Application of active contours for photochromic tracer flow extraction," *IEEE Transactions of Medical Imaging*, vol. 16, no. 3, pp. 284–293, June 1997.

[9] A.K. Klein, L. Forester, and A.A. Amini, "Quantitative coronary angiography with deformable spline models," *IEEE Transactions of Medical Imaging*, vol. 16, no. 5, pp. 468–482, October 1997.

[10] A. Yezzi, S. Kichenassamy, A. Kumar, R. Olver, and A. Tannenbaum, "A geometric snake model for segmentation of medical imagery," *IEEE Transactions of Medical Imaging*, vol. 16, no. 2, pp. 199–209, April 1997.

[11] M. Stone and L. Lundberg, "Three-dimensional tongue surface shapes of english consonants and vowels.," *The Journal of The Acoustical Society of America*, vol. 99, no. 6, pp. 1–10, 1996.

[12] M. Unser and M. Stone, "Automatic detection of the tongue surface in sequences of ultrasound images," *The Journal of The Acoustical Society of America*, vol. 91, no. 5, pp. 3001–3007, May 1992.

[13] Yves Laprie and Marie-Odile Berger, "Extraction of tongue contours in x-ray images with minimal user interaction," in *Fourth International Conference on Spoken Language Processing*, 1996.

[14] Michael A. Gennert and Alan Y. Yuille, "Determining the optimal weights in multiple objective function optimization," in *Proceedings of Second International Conference on Computer Vision*, 1988, pp. 87–89.

[15] K. F. Lai and R. T. Chin, "Deformable contours- modeling and extraction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 11, pp. 1084–1090, 1995.

[16] V.S. Nalwa, *A Guided Tour of Computer Vision*, Addison-

TABLE II
MSSD's for the two sequence of Figure 15 and Figure 16.

Typical Sequence				Difficult Sequence			
Frame No	System vs Manual1	System vs Manual2	Manual1 vs Manual2	Frame No	System vs Manual1	System vs Manual2	Manual vs Manual2
1	3.0097	1.8238	3.1504	1	5.0630	5.2169	2.2422
2	2.9019	1.9735	2.9639	2	3.2623	3.9767	3.0458
3	3.3885	3.0704	2.0937	3	4.0115	3.9238	1.5034
4	3.0894	1.6269	2.6472	4	5.8412	5.3858	3.6656
5	2.4469	1.5346	2.3724	5	6.0475	3.6786	5.7572
6	5.4399	4.7912	1.8579	6	4.3225	3.8778	3.7938
7	2.9381	1.6918	1.9776	7	3.6290	3.1380	3.0022
8	2.4045	1.6886	2.2610	8	3.1220	3.4073	4.4441
9	2.8251	1.3422	2.2909	9	3.0558	4.4931	4.6220
10	2.3821	1.3573	2.4711	10	3.5803	4.5661	3.1800
11	2.1171	1.6582	2.2988	11	4.1133	2.6293	5.4515
12	2.0572	2.4057	2.0397	12	2.8770	5.1898	3.3192
13	2.1382	1.1056	2.1424	13	3.7125	4.9760	4.4436
14	3.1144	1.7350	3.4523	14	6.2063	8.7588	5.8457
15	2.4409	1.9860	1.7880	15	6.1537	8.5201	8.4916

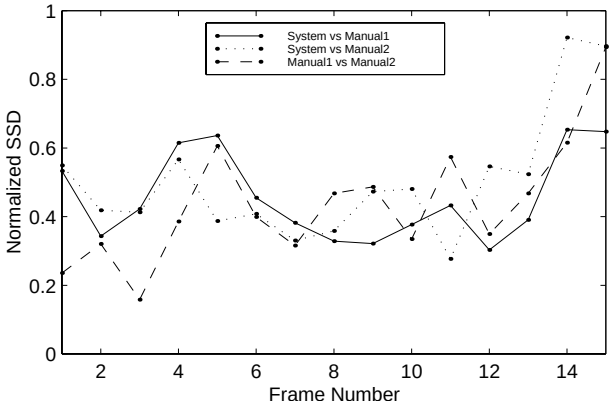


Fig. 15. The NMSSD values for a difficult test case.

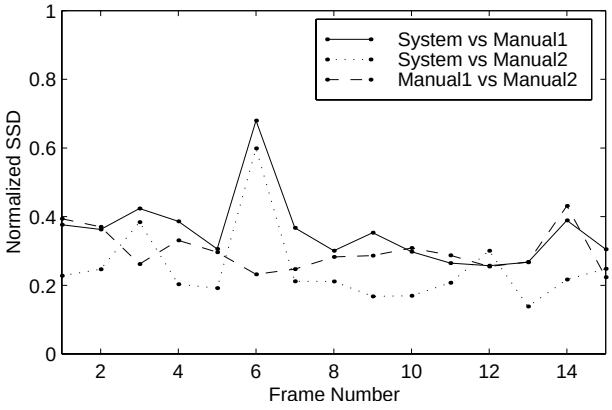


Fig. 16. The NMSSD values for a typical test case.

Wesley, 1993.

[17] M.O. Berger, N. Maurice, G. Winterfeldt, and J.P. Lethor, "Automatic 3D reconstruction of the beating left ventricle using transthoracic echographic images," *Computers in Cardiology*, vol. 25, pp. 641–644, 1998.

[18] V. Chalana, D.T. Linker, D.R. Haynor, and Y. Kim, "A multiple active contour model for cardiac boundary detection on echocardiographical sequences," *IEEE Transactions of Medical Imaging*, vol. 15, no. 3, pp. 290–298, 1996.

[19] A. Blake and M. Isard, *Active Contours*, Springer-Verlag, 1998.

[20] S. Malassiotis and M. Gerassimos, "Tracking the left ventricle in echocardiographic images by learning heart dynamics," *IEEE Transactions of Medical Imaging*, vol. 18, no. 3, pp. 282–290, 1999.

[21] D. Terzopoulos, A.P. Witkin, and M. Kass, "Constraints on deformable models: Recovering 3d shape and nonrigid motion," *Artificial Intelligence*, vol. 36, no. 1, pp. 91–123, 1988.

[22] L. D. Cohen, "Note: On active contour models and balloons," *CVGIP-Image Understanding*, vol. 53, no. 2, pp. 211–218, 1991.

[23] A. A. Amini, T.E. Weymouth, and R.C. Jain, "Using dynamic programming for solving variational problems in vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 9, pp. 855–867, 1990.

[24] 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*, vol. 17, pp. 294–302, 1995.

[25] D. J. Williams and M. Shah, "A fast algorithm for active contours and curvature estimation," *Computer Vision, Graphics, Image Processing*, vol. 55, pp. 14–26, 1992.

[26] Steve R. Gunn and Mark S. Nixon, "A robust snake implementation; a dual active contour," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 1, pp. 63–68, January 1997.

[27] S. Singh and S.S. Singh, *Phonetics Principles and Practices*, University Park Press, 1976.

[28] D.R. Calvert, *Descriptive Phonetics*, Brian C. Decker, 1980.

[29] V. Chalana and Y. Kim, "A methodology for evaluation of boundary detection algorithms on medical images," *IEEE Transactions of Medical Imaging*, vol. 16, no. 6, pp. 642–652, 1997.