Robust tongue tracking in ultrasound images: a multi-hypothesis approach

Catherine Laporte¹, Lucie Ménard²

¹Department of Electrical Engineering, École de technologie supérieure, Montreal CANADA
²Department of Linguistics, Université du Québec à Montréal, Montreal CANADA
catherine.laporte@etsmtl.ca menard.lucie@uqam.ca

Abstract
Ultrasound (US) imaging is an excellent means of observing tongue motion during speech. Tracking the tongue contour in US video is required for analysis of this motion, but most currently available techniques suffer from either a lack of temporal consistency or a lack of robustness to difficult conditions such as a rapidly deforming tongue or momentarily poor image quality. This paper proposes a new algorithm combining active contours, active shape models and particle filtering that addresses these shortcomings. The strength of this approach lies in the fact that it maintains multiple tongue shape hypotheses simultaneously. Experimental results show that this approach outperforms a classic active contour algorithm as well as a shape-constrained variant thereof, particularly in difficult tracking conditions.

Index Terms: tongue tracking, tongue segmentation, ultrasound, particle filter, active shape model

1. Introduction
The tongue is possibly the most important articulator involved in speech production, and it is also one whose motion, unlike that of the lips, is mostly hidden from view in normal conversation. Observing and characterizing tongue motion are of interest to the study of speech production itself, as well as a promising avenue toward the development of effective strategies for speech therapy. Ultrasound (US) imaging lends itself extremely well to this purpose as it provides a dense representation of the moving tongue (typically a full 2D sagittal slice) and is also non invasive, portable and inexpensive [1].

Analysis of tongue shape and motion from US images, as performed in typical speech science studies, requires that the tongue contour be segmented in a large number of video frames. This task is extremely time consuming unless it is automated at least in part. The literature proposes many methods for segmenting the tongue in US videos. The oldest and best known methods model the tongue surface as an active contour (or snake) [2] attracted toward areas of strong image gradient and tongue-like features (such as a bright white band [3]) with constraints on the amount of bending and stretching [4]. This is the approach implemented in the free and widely used Edgetrack software [3], where optimized snakes are copied from one frame to the next to be used as initial guesses. One shortcoming of snakes is that they can produce contour shapes that are quite uncharacteristic of a tongue during speech. One way to address this is to enforce global constraints using a statistical shape model, as proposed by Roussos et al. [5]. Their approach builds the shape component of an active appearance model [6] using a database of X-ray images where the entire vocal tract is visible (which is not the case for US) and the texture component of the active appearance model from US data. Similarly, in recent work [7], we trained an active shape model (ASM) [8] exclusively from US data (thereby avoiding the need for invasive X-ray imaging) and used it in combination with snakes in an iterative framework inspired by the work of Hamarneh et al. [9]. If a database of segmented training data is available, another possibility is to estimate a direct mathematical relationship between raw US image data and the vertices of the correct tongue contour using neural networks [10].

All these approaches suffer from weak temporal consistency: the progression of tongue shape from one frame to the next is not well modeled. At one extreme, current snake-based and statistical shape model-based approaches use only the tongue contour identified in the previous frame to predict the one in the current frame, which often fails in the case of rapid tongue deformation. At the other extreme, the neural network approach [10] treats each frame independently, thereby under-constraining possible tongue motions from one frame to the next. Tang et al. [11] address some of these problems by modeling the evolution of tongue contours as a graph in the 2D+t space of the video sequence, thereby elegantly expressing the segmentation problem as a global optimization over the graph. This allows future frames to condition the current tongue contour, imposes global deformation continuity constraints over time and improves the segmentation, particularly in cases of sudden motion or momentarily poor image quality. However, this approach still fails to exploit available knowledge and data to explicitly model likely speech-induced tongue deformations.

In contrast, this paper demonstrates that using an explicit statistical model of tongue shape and motion and maintaining multiple hypotheses during tracking also greatly improves tracking robustness. A particle filtering approach [12] is proposed to track the tongue in US video sequences. For this purpose, tongue shape is summarized using an ASM [8], and multiple tongue deformation hypotheses are sampled from a statistical model describing the evolution of shape parameters over time, as learned from a database of segmented videos. In each video frame, the most likely tongue shape hypothesis is refined using a snake [3].

The remainder of this paper is structured as follows. Section 2 describes the shape and motion models used for tracking and provides the details of the proposed tracking algorithm. Section 3 presents experiments showing how the proposed approach offers superior tongue segmentation performance over a traditional snake-based approach and our previous approach combining snakes and ASMs. Section 4 summarizes our contribution and outlines directions for future work.

2. Proposed approach
Figure 1 illustrates the proposed tracking approach. Put simply, multiple plausible tongue contour hypotheses, called parti-
To model likely evolutions of tongue contour state over time, a simple state transition model is estimated from the same database that was used to build the ASM. The evolution of $x$, $y$, $s$, $w_1$, $w_2$ and $w_3$ over time is studied by considering pairs of consecutive images within each of the US video sequences in the training database. A compact state vector is associated with each of the contours in the training images and the differences between consecutive state vectors are computed. These are pooled together to compute a $6 \times 6$ covariance matrix $\Sigma$ expressing the distribution of likely evolutions in tongue state from one frame to the next. Assuming that the average tongue deformation from one frame to the next is zero, this provides a multivariate Gaussian state transition model that can predict a variety of possible tongue states (including their position, scale and shape) in one frame based on its state in the previous frame.

2.2. Tracking the tongue

To track the state of the tongue in new US video sequences, we use the particle filter illustrated in Fig. 1. The initial tongue contour is first associated with a compact state vector, as described in section 2.1. $N$ copies of this state vector, called particles, are then generated and plausible tongue states for the next frame are randomly sampled for each of these particles using the proposed multivariate Gaussian state transition model. The likelihood of each state is then assessed based on the image data encountered in the next frame. Specifically, the state of each particle is transformed back into a set of tongue contour vertices, forming a snake whose configuration (i.e. the positions of its vertices $v_i$, $i = 1, \ldots, n$) is then optimized to latch to local gradient features in the image according to the energy functional

$$E_{\text{snake}} = \sum_{i=1}^{n} \alpha E_{\text{int}}(v_i) + \beta E_{\text{gradient}}(v_i).$$

In (1), $E_{\text{int}}$ is the local internal energy of the snake, encoding local curvature and stiffness constraints such that

$$E_{\text{int}}(v_i) = \lambda_1 \left(1 - \frac{||\vec{v}_{i-1}v_i||}{||\vec{v}_{i-1}v_{i+1}||} \right) + \lambda_2 \frac{||\vec{v}_{i-1}v_i|| - d}{d},$$

where $d$ is the average distance between two consecutive vertices. $E_{\text{gradient}}(v_i)$ moves the contour towards regions of high image gradient (see Li et al.’s paper for details [3]). We empirically set the snake’s parameters as follows: $\alpha = 0.8$, $\beta = 0.2$, $\lambda_1 = 0.95$, $\lambda_2 = 0.05$. The energy functional of (1) can be optimized accurately and quickly, but is not an ideal measure of tongue contour quality since US speckle sometimes creates areas with high image gradient that are unrelated to tongue contour. Therefore, to evaluate the quality of each optimized contour, we use the more complete (but slower to compute) energy functional proposed by Li et al. [3]:

$$E'_{\text{snake}} = \sum_{i=1}^{n} \alpha E_{\text{int}}(v_i) + \beta E_{\text{gradient}}(v_i) E_{\text{band}}(v_i),$$

where $E_{\text{band}}(v_i)$, measures the contrast between the region immediately above the contour (which should be a bright band) and the region immediately below it. The reader is once again referred to Li et al.’s paper [3] for the mathematical definition of $E_{\text{band}}$. Using (1) for particle optimization and (3) for particle evaluation provides a reasonable compromise between accuracy and computation time. This is important because the
proposed approach must optimize a contour and evaluate it \( N \) times (once for each particle) in each frame.

The energy of the snake, as given in (3) measures the likelihood that each particle accurately represents the true state (position, scale and shape) of the tongue. The relationship between energy and likelihood is defined as

\[
L = \exp(-E_{\text{snake}}),
\]

which provides numbers lying on the closed interval \([0, 1]\).

Each particle, \( i \), \( i = 1, \ldots, N \) is then assigned a weight \( W_i \), \( 0 \leq W_i \leq 1 \) based on its likelihood \( L_i \):

\[
W_i = \frac{L_i}{\sum_{j=1}^{N} L_j}
\]

The particle with the largest weight is considered the best solution to the tongue segmentation problem for the current frame, and the corresponding snake is further refined by optimizing the full energy functional (3). The other particles also represent plausible hypotheses and are not forgotten: \( N \) new particles are sampled from the current set with replacement, with each particle having a probability of being chosen proportional to \( W_i \).

This maintains a set of plausible segmentation hypotheses from which segmentation in future frames can benefit.

To further control computation time, the number of particles \( N \) is adjusted at every frame according to the uncertainty in tongue state. This is done during the evaluation of particle likelihood, as proposed by Fox et al. [13]: particles are evaluated (and \( N \) is increased) until the total likelihood \( \sum_{j=1}^{N} L_j \) reaches an empirically devised threshold of three times the likelihood of the first, manually initialized contour. If there is a lot of uncertainty (e.g., image quality is very poor for a few frames), \( N \) will increase to meet this threshold. \( N \) is also not allowed to shrink below 50 or grow beyond 300. These numbers were chosen empirically as a reasonable compromise between accuracy and computation time. The algorithm was not found to be very sensitive to these choices.

3. Experiments

3.1. Data acquisition and model estimation

The images used in this study were acquired using a Sonosite 180 plus US scanner with a micro-convex 8-5 MHz transducer set at a 94 degree field of view and a depth of 8.9 cm (adults) or 7.4 cm (children). During each recording, the subject’s head was kept immobile using a helmet bolted to the wall and the US transducer was kept immobile using a microphone stand.

The proposed approach was evaluated experimentally and compared to our implementation of Li et al.’s snake-based approach [3] and our previous approach based on the combination of snakes and ASMs [7]. For this purpose, an ASM and a statistical motion model were first estimated from a training database containing 48 video sequences, for a total of 1978 tongue US images. These video sequences show the moving tongue of two different healthy adults uttering \([V_pV]\) and \([V_tV]\) tongue US images. These video sequences were obtained using snake-based segmentation and interactive correction in the same fashion as for the segmentation of the training database. For fairness, the three tracking algorithms were launched from the same manually initialized contour.

3.2. Segmentation quality evaluation

Segmentation quality was evaluated by evaluating the mean sum of distances (MSD) between the ground truth tongue contours and their corresponding segmentation results from the three tracking algorithms. The MSD between two contours (i.e., lists of \( n \) vertices) \( u \) and \( v \), proposed by Li et al. [3], is given by

\[
\text{MSD}(u, v) = \frac{\sum_{i=1}^{n} \min_j \| \mathbf{v}_i - \mathbf{u}_j \| + \sum_{v_i} \min_j \| \mathbf{v}_i - \mathbf{u}_j \|}{2n}
\]

The local contour energy, normalized with respect to the local contour energy of the first frame in each sequence (which is assumed to be correctly segmented since initialization is done manually), was also evaluated for each algorithm. This is computed for each vertex \( v_i \) and each frame \( j \) as

\[
E_j^i(v_i) = \frac{|E_j^i(v_i) - E_{\text{init}}^i(v_i)| \times 100%}{E_{\text{init}}^i(v_i)}
\]

where the subscript ‘init’ refers to the (manually initialized) first contour of the sequence considered. This is precisely the quality measure that was exploited during ground truth segmentation, to provide the operator with visual indications of tracking errors in the snake-based result [14]. The normalized energy measure is thus also available for the ground truth segmentation and is of particular relevance to the analysis of our results. In principle, the ground truth tongue contours should have low normalized energy, unless the image quality was poor, or the tongue adopted a particularly convoluted shape, or there remained some inaccuracies in the segmentation due to human factors, all of which occurred in our experiments to some extent. Thus, the normalized energy of the ground truth contours provides an indication of the degree of challenge posed by the particular video sequence considered.

3.3. Results

Figure 3 (top) shows a box plot of the MSD measures, evaluated for every frame, grouped by video sequence and by segmentation method. The basic snake method (‘snake’) yielded the worst segmentation accuracy among the approaches tested. Constraining the snake using an ASM (‘asm’) greatly improved tracking accuracy in most cases, while maintaining multiple contour hypotheses using the particle filter (‘pf’) generally resulted in further improvements. This is apparent not only in the median of the MSD distributions but also in their range, which reflects the ability of the algorithms to quickly recover from segmentation errors.
Figure 3: Distribution of segmentation quality measures (top: MSD to ground truth, bottom: normalized contour energy) for 9 video sequences. `snake`: snake-based segmentation, `asm`: shape constrained snake segmentation, `pf`: particle filter segmentation, `truth`: ground truth segmentation. Dots: median, boxes: interquartile range, whiskers: extremes, circles: outliers.

Figure 3 (bottom) shows a box plot of the normalized contour energy (averaged over contour vertices), evaluated for each frame, grouped by video sequence and by segmentation method, including ground truth segmentation. These data (most notably their interquartile ranges) support the conclusions drawn from the MSD results, showing `pf` to generally outperform the other algorithms. Interestingly, `pf` often yielded lower normalized energy than the semi-manually obtained ground truth contours. This suggests that the human interactions involved in ground truth segmentation sometimes led to slightly sub-optimal contours, which the particle filter actually improved upon.

Figure 4 presents detailed results for two video sequences which posed particular challenges. In the first example (sequence s4, Figure 4(a)), the subject swallowed between two [VpV] utterances. The long bright regions in the normalized energy maps (displayed in the leftmost column) show that the `snake` and `asm` algorithms lost track of the tongue early in the sequence (when swallowing began) and were never able to fully recover. However, the particle filter was able to generate plausible hypotheses in the vicinity of the true tongue state a few frames within the swallow, leading to a graceful recovery.

In the second example (sequence s7, Figure 4(b)), several tongue contours were missing due to loss of contact with the transducer during recording. The evolution of the segmented tongue contour across this part of the sequence is highly characteristic of the different tracking algorithms tested. The `snake` algorithm essentially copied the last contour it had optimized without shape constraints, respectively. Future work will explore two main directions: (1) improving the computational efficiency of the particle filter by further simplifying the particle evaluation step and using alternative contour quality measures (possibly some of the ones proposed by Tang et al. [11]) and (2) automatically detecting tracking errors and recoveries so as to perform retroactive corrections on incorrectly segmented frames, thereby possibly providing a faster alternative to techniques based on global optimization over entire video sequences [11].

4. Conclusions

This paper proposed a new algorithm based on the combination of snakes, ASM's and a particle filter to segment and track the tongue in US video sequences acquired during speech. The algorithm’s strength comes from its ability to maintain several tongue contour hypotheses, which enables swift recovery from occasional tracking errors, particularly in the presence of rapid, high curvature tongue deformations or following a temporary loss in image quality. Experimentally, this approach compared favorably to two other approaches based on snakes, with and without shape constraints, respectively. Future work will explore two main directions: (1) improving the computational efficiency of the particle filter by further simplifying the particle evaluation step and using alternative contour quality measures (possibly some of the ones proposed by Tang et al. [11]) and (2) automatically detecting tracking errors and recoveries so as to perform retroactive corrections on incorrectly segmented frames, thereby possibly providing a faster alternative to techniques based on global optimization over entire video sequences [11].
5. References


