Automatic Tongue Contour Tracking in Ultrasound Sequences without Manual Initialization

Hongcui Wang[1], Siyu Wang[1], Bruce Denby[1,2], Jianwu Dang[1,3,*]
[1]Tianjin University, Tianjin, China
E-mail: hcwang@tju.edu.cn, syuwang@tju.edu.cn
[2]Université Pierre et Marie Curie, Paris, France
E-mail: denby@ieee.org
[3]Japan Advanced Institute of Science and Technology, Ishikawa, Japan
E-mail: jdang@jaist.ac.jp

Abstract—**Tracking the movement of the tongue is important for understanding how tongue shape change contributes to speech production and control. Ultrasound imaging is widely used to record real time information on the tongue surface; however, noise, artefacts, and the presence of spurious edges render automatic detection of tongue contours without manual initialization difficult. In this paper, we propose a method to extract ultrasound tongue surface contour in a totally automatic way using a three-step procedure: 1) noise reduction using a non-local mean filter; 2) use of a quadratic function to roughly fit the surface contour based on points obtained with a Robert cross operator; and 3) an automatic refinement based on gradient shift and relative distance of candidate points to the initial rough contour point. Experiments are conducted on isolated vowels and on a continuous utterance of vowel sequence. The Mean Sum of Distances criterion shows that the proposed method provides results on a par with the popular EdgeTrak algorithm on these two data sets, as compared to hand-scanned contours, but without any manual initialization.**

I. INTRODUCTION

Speech production is often modeled as a source-filter [1] system, with vocal tract shape, in large part due to tongue motion, playing the role of a time-varying acoustic filter. Studies involving a variety of techniques have been carried out to study vocal tract shape, such as X-ray[2], magnetic resonance imaging (MRI) [3][4], electromagnetic articulography (EMA) [5][6], and ultrasound imaging [7][8], where each modality has its own advantages and drawbacks. Recently, many researchers have taken to combining different modalities in order to obtain a more complete analysis of vocal tract shape [9][10][11].

Tongue shape changes contribute in a primordial way to speech production and control, and understanding the mechanism of the tongue movement can also increase our knowledge of speech disorders. Our laboratory is developing an automatic system to model tongue surface shape changes during speaking by fusing ultrasound and EMA data, as illustrated in figure 1. As the first step, we wish to efficiently track the tongue surface shape movement based on real time ultrasound images.

Fig. 1 Data collection system showing ultrasound and EMA devices

A variety of algorithms have been proposed to deal with the tongue contour tracking problem. A well-known example is snakes [12], which make use of the "energy", derived from the image gradient. Such gradients are typically noisy but constraints based on homogeneity of intensity allow contours to be successfully extracted [9], as in for example the popular EdgeTrak [13] software. EdgeTrak however, requires 5-6 points in the first tongue image to be entered by hand; it also must be reinitialized periodically when found contours begin to drift over time.

In this paper, we propose a simple but efficient method to extract tongue surface contours automatically without any manually input points. The algorithm first uses a non-local mean filter to perform image denoising; then makes use of a Robert cross operator to obtain candidate points on the tongue, to which a quadratic fit is made in order to get a first rough tongue surface shape. As a final step, contour point positions are automatically refined using gradient shift and relative distance measures. The algorithm is detailed in the following

* the corresponding author

section. Results on experimental data are presented in section III, while a conclusion appears in section IV.

II. TONGUE TRACKING METHOD

The principal idea behind the method is the combination of an initial rough fit followed by an adjustment based on gradient shift and relative distance. Before describing this procedure, the preliminary image denoising step is first presented.

A. Image denoising

In this step stochastic noise is reduced by averaging over groups of images; subsequently, speckle reduction is performed with a non-local mean filter.

1) Averaging

We average each frame with its previous and subsequent frames as in Equation (1):

$$g(m,i)=\frac{1}{3}\sum_{q=-1}^{1} g(m,i+q) \tag{1}$$

where $g(m,i)$ is the grey value of m^{th} pixel number in i^{th} frame and q is an integer. This operation can reduce some noise and at the same time make the boundary more clearly.

2) Speckle reduction with a non-local mean filter

As we all known, the common noise model in ultrasound imaging is speckling noise, some algorithms are proposed to reduce it [14]. Here, we use the non-local means filter, which calculates the mean of the pixels in an image after weighting them by their degree of similarity to the target pixel, which is known to help preserve detail in filtered images [15]. For an image of area Ω described by pixel coordinates p and q, the algorithm is:

$$u(p)=\frac{1}{C(p)}\int_{\Omega} v(q)\, f(p,q)\, dq \tag{2}$$

where $u(p)$ is the value of the filtered image at the point p, $v(q)$ the unfiltered value at q, and $f(p,q)$ the weighting function that determines the degree of similarity of the image at points p and q. The normalizing factor $C(p)$ is given by:

$$C(p)=\int_{\Omega} f(p,q)\, dq \tag{3}$$

The function $f(p,q)$, can take many forms. A common choice is the Gaussian weighting function given by:

$$f(p,q)=e^{-\frac{|B(q)-B(p)|^2}{h^2}} \tag{4}$$

where the standard deviation h is the filter parameter and $B(p)$ the local mean of image points in a 4×4 window centered on p.

B. Tongue contour extraction

In this part, we first make a "rough" fit of a convex curve to the points obtained from the Robert cross operator, and then modify the curve shape based on gradient shifts.

1) Rough curve fit

As a first step, we obtain candidate points for the lower edge of the tongue surface in Ultrasound images using the simple and well-known Roberts cross edge-detection operator [16].

A fit to the candidate points is performed with a quadratic function, which is sufficient since a physical tongue cannot assume arbitrarily complex shapes.

In order to help deal with outliers, adjacent frame information is used, on the assumption that a specific tongue point will not move far in the interval between two frames. Specifically, we average the contour of frame i-1 and i+1 to obtain the contour of frame i.

2) Contour refinement

To improve the fit, we measure the gradient at points above and below each candidate point, after multiplying by a factor that down-weights point far from the initial curve. The point (x, y) with the highest gradient is then chosen as the new candidate. The weighting factor is modeled as a Gaussian distribution in the relative distance (y-y'), as follows:

$$G(x,y)=\left[I(x,y+1)-I(x,y)\right]\times\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y-y'-\mu)^2}{2\sigma^2}\right) \tag{5}$$

where $I(x, y)$ is the grey value of the point in the image, and the $[I(x, y+1)-I(x, y)]$ is the original gradient of the candidate point (x, y). Empirically, the values μ=0, σ=9~11 were found to give the best results.

At last, cumulative moving average strategy is used to do the smoothing for the whole frames. The moving average filter window is set 20 in our experiments.

III. EXPERIMENTS AND RESULTS

A. Data

There are two types of experimental Data, four single vowels 'a', 'e', 'i', 'o' read by a male speaker, and a continuous sequence of all vowels 'aoeiuv', read by a female speaker, with each set read 5 times. The Ultrasound system used is a Terason T3000 with the 8MC3 probe [17].

B. Preprocessing

Initial frames before and after the utterance are first removed. As an example, for the vowel /a/, 52 utterance images remain from an initial 187 after this cleaning pass.

C. Measurement

In order to verify our results quantitatively, we compare the difference between the automatic tracking results and the manual contours drawn by speech scientists. The difference between two contours was calculated using a Mean Sum of Distance (MSD) [18]. The MSD between two contours $U=\left[u_1,u_2,\cdots,u_n\right]$ and $V=\left[v_1,v_2,\cdots,v_n\right]$ is defined as

$$MSD(U,V)=\frac{1}{2n}\left(\sum_{i=1}^{n}\min_j\left|v_i-u_j\right|+\sum_{i=1}^{n}\min_j\left|u_i-v_j\right|\right) \tag{6}$$

D. Results

Comparison was also made with EdgeTrak, as shown in Table 1. For Edgetrak, we input the original ultrasound images of one utterance, e.g. 52 frames for utterance /a/, and

manually label six points uniformly along the tongue surface, only on the first image, even for the case of the continuous sequence of vowels. The mean and standard derivation of the distance errors between the tracking results and the manual contours drawn by the speech scientists are listed in Table 1. In the experiment, the speech scientists labeled 30 points on the contour for each frame.

Table I
MEAN/SD DISTANCE ERRORS IN PIXELS

MSD (mean/SD)	EdgeTrak		Proposed Method	
'a'	4.7	1.9	**2.7**	1.4
'e'	5.8	2.2	**5.7**	2.2
'i'	8.6	4.9	**2.7**	0.8
'o'	4.4	2.2	**4.1**	1.3
'aoeiuv'	1.8	0.9	**1.7**	0.7

One pixel here is 0.295mm. As the numbers in bold in Table 1 indicate that, the automatic contours by our proposed method are in rather good agreement with the expert-drawn contours, and wholly competitive with EdgeTrak. Indeed it is known that EdgeTrak contours have a tendency to drift over time, requiring re-initialization. Our results, furthermore, are obtained without recourse to initialization, whereas 6 points are hand-labeled on the first frame for EdgeTrak. The following figures illustrate results. The tracking results of the two methods for Figure 2 are shown in Figure 3. Another ultrasound image sequence is in Figure 4, and its tracking results for the two methods appear in Figure 5. Notice that we did not re-initialize when the tendency drift when using EdgeTrak system.

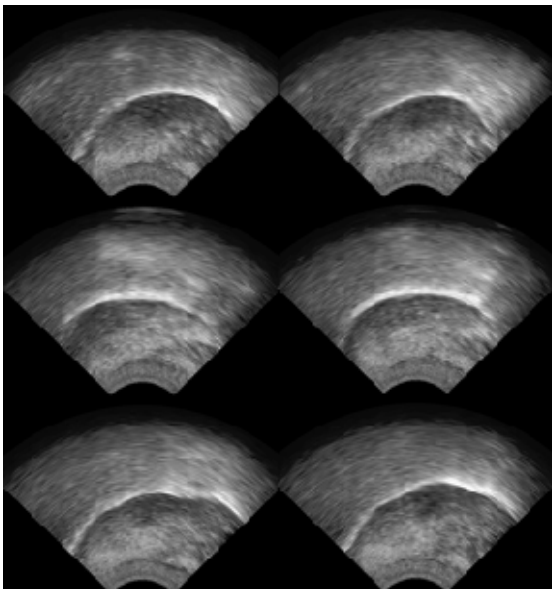

Fig. 2 Image sequence of the tongue during vowel /a/. Every 10th frame from 52 frames is shown. Images are ordered from top to bottom, left to right.

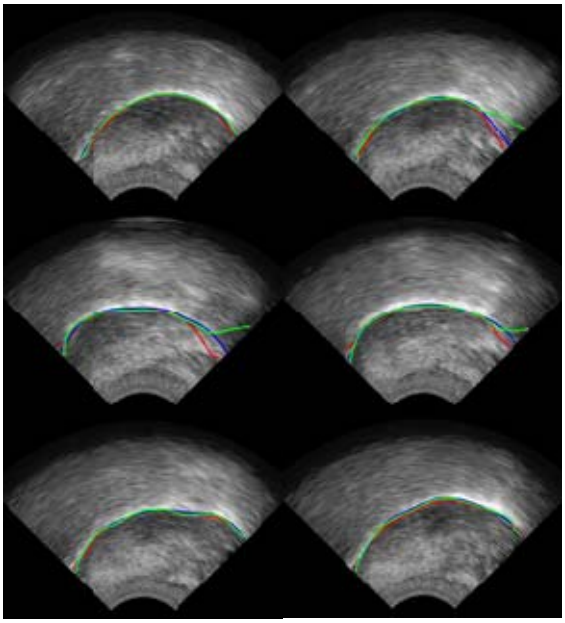

Fig. 3 Tracked contours for the sequence in Fig.2. The red curve is the result of the proposed method. The blue curve is the curve manually tracing. The green curve is the result of Edgetrak.

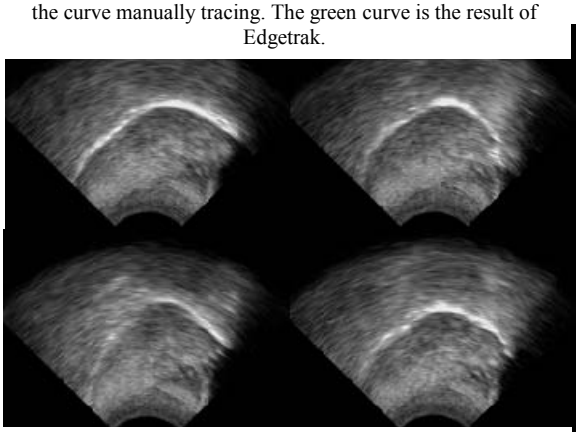

Fig. 4 Image sequence of the tongue during vowel /i/. Every 10th frame from 36 frames is shown. Images are ordered from top to bottom, left to right.

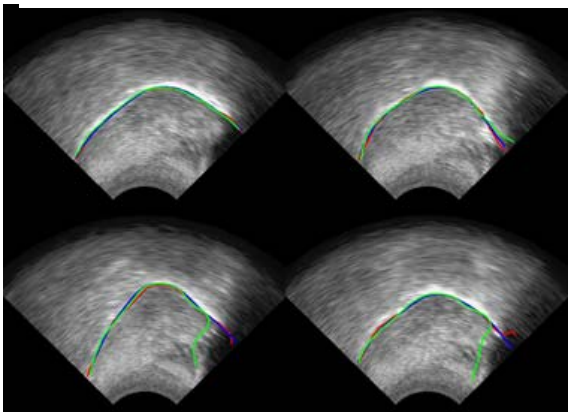

Fig. 5 Tracked contours for the sequence in Fig.4. The red curve is the result of the proposed method. The blue curve is the curve manually tracing. The green curve is the result of Edgetrak without re-initialization while drifts occur.

Figure 3 shows the tracked contours of /a/ for the both methods are similar and close to the curve manually tracing. But for some Ultrasound images, the results of two methods are different. We can see from Figure 5 that the results of our method are relatively stable and close to the curve manually tracing, even from the first few frames without any initialization. The tracked contours of the Edgetrak begin to drift over ten frames and may become worse over time.

IV. CONCLUSIONS

We propose a simple but effective method to extract tongue surface contour from the ultrasound images, in which a rough curve fitting with quadratic function followed by a refinement based on gradient shifts are carried out. The robustness and the effectiveness of the proposed method have been verified by comparing the automatic tracking results both with the manual contours drawn by speech scientists and EdgeTrak. Unlike EdgeTrak system, we do not require the user to input initial starting points along the tongue surface. Experimental results show that our proposed method agrees well with the expert contours and is competitive with EdgeTrak, without any initialization.

ACKNOWLEDGMENT

The research is partly supported by the National Basic Research Program of China (No. 2013CB329301), the National Natural Science Foundation of China (No. 61303109), and the PhD Programs Foundation of Ministry of Education of China (No. 20120032120043).

REFERENCES

[1] G. Fant, Acoustic theory of speech production, The Hague: Mouton, 1960.

[2] P. Ladefoged, R. Harshman, L. Goldstein, and L. Rice, "Generating vocal tract shapes from formant frequencies," *The Journal of the Acoustical Society of America*, vol. 64, no. 4, pp. 1027–1035, 1978.

[3] T. Baer, and J. C. Gore et al., "Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels," *The Journal of the Acoustical Society of America* , vol. 90, no. 2, pp. 799-828, 1991.

[4] K. Honda, H. Takemoto, T. Kitamura, S. Fujita, and S. Takano, "Exploring human speech production mechanisms by MRI," *IEICE Transactions on Information and Systems*, vol.87, no.5, pp.1050-1058,2004.

[5] J. Perkell, M. Cohen, M. Svirsky, M. Matthies, I. Garabieta, and M. Jackson, "Electro-magnetic mid-sagittal articulometer (EMMA) systems for transducing speech articulatory movements," *The Journal of the Acoustical Society of America*, pp. 3078-3096, 1992.

[6] X. Lu and J. Dang, "Vowel production manifold: Intrinsic factor analysis of vowel articulation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp.1053-1062, 2010.

[7] K. L. Watkin and J. M. Rubin, "Pseudo-three-dimensional reconstruction of ultrasonic images of the tongue," *The Journal of the Acoustical Society of America*, vol. 85, no. 1, pp. 496–499, 1989.

[8] E. Slud, P. Smith, M. Stone, and M. Goldstein, "Principal components representation of the two-dimensional coronal tongue surface," *Phonetica*, vol. 59, No. 2-3, pp. 108-133, 2002.

[9] M. Stone, V. Parthasarathy, K. Iskarous, M. NessAiver, and J. Prince, "Tissue strains and tongue shapes: combining tMRI and ultrasound," *In Proceedings of the Fifteenth International Congress of Phonetic Sciences*, pp. 3-9, 2003.

[10] A. Katsamanis, G. Papandreou, P. Maragos, "Face active appearance modeling and speech acoustic information to recover articulation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 3, pp. 411-422, 2009.

[11] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, "Silent speech interfaces," *Speech Communication,* vol. 52, no. 4, pp. 270-287, 2010.

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

[13] M. Li, C. Kambhamettu, and M. Stone, "Snake for band edge extraction and its applications," *In Computer Graphics and Imaging*, pp. 261-266, 2003.

[14] P. Coupé, P. Hellier, C. Kervrann, and C. Barillot, "NonLocal means-based speckle filtering for ultrasound images," *IEEE Trans. Image Process.*, vol. 18, no. 10, pp. 2221-2229, 2009.

[15] A. Buades, B. Coll, and J. M. Morel, "A non-local algorithm for image denoising," *In Computer Vision and Pattern Recognition*, Vol. 2, pp. 60-65, 2005.

[16] L.G. Roberts, "Machine perception of three-dimensional solids," *MIT Lincoln Laboratory Technical Report*,1965.

[17] http://www.terason.com/t3000/, Terason Corporation, 77 Terrace Hall Ave, Burlington, MA 01803, USA

[18] M. Li, C. Kambhamettu, and M. Stone, "Automatic contour tracking in ultrasound images," *Clinical Linguistics & Phonetics*, Vol.19 no. 6-7, pp. 545–554, 2005.