Speckle Tracking for the Recovery of Displacement and Velocity Information from Sequences of Ultrasound Images of the Tongue

Mathews Jacob, Heike Lehnert-LeHouillier, Sourabh Bora, Stephen McAleavey, Diane Dalecki, and Joyce McDonough

University of Rochester
E-mail: mathews.jacob@rochester.edu

Abstract

This paper explores the use of deformable registration (speckle tracking) as a method for obtaining point correspondences on sequences of tongue images acquired via ultrasound in order to estimate tongue velocity and displacement. We model the velocity field as smooth Bspline functions and estimate it from adjacent frames. This model enables us to accumulate the motion and hence calculate the displacement from the first frame to any frame in the image sequence. Specifically, it provides the displacement estimates on the curve, enabling us to introduce "virtual flesh point markers" on the tongue surface. We tested the algorithm on ultrasound image sequences that were taken during the production of domain initial vowels in two different prosodic domains – word initial and accentual phrase initial. The results demonstrate the utility of the algorithm in quantification of tongue motion in speech.

1 Introduction

The main goal of the use of imaging techniques in speech research is to improve our understanding of the configuration of the articulators as well as the change of the configuration of the articulators over time during speech. Ultrasound is increasingly used as an imaging tool for speech research, and has proven successful in imaging the configuration of the tongue with respect to the hard palate [9], and [11].

The classical approach to quantify tongue motion is by extracting the contours of the tongue surface from ultrasound using active contours [4]. However, this approach can only measure the shape of the tongue. In contrast to imaging methods such as x-ray microbeam and electromagnetic articulography that track flesh points on the tongue, the above mentioned ultrasound technique has limited capability to provide point correspondences. This makes it difficult to quantitatively compare or average the contours taken at two different points of time. Various curve post-processing approaches to address this problem were introduced in [6] and [8]. However, the success of these methods in deriving corresponding points were limited due to the drastic change in length and shape of the tongue during speech. Moreover, these algorithms only exploit the information at the tongue surface and hence are not robust to faint contours (due to the tongue surface being orthogonal to the transducer) and exclusions.

The main objective of the current paper is to explore speckle tracking as a method for obtaining point correspondences on sequences of ultrasound images of the tongue. Although speckle is often considered as an artifact or noise in conventional ultrasound imaging, it provides a signature for precise motion tracking. The acoustic scatterers producing the speckle patterns displace with the tissue, and hence lead to moving speckle patterns that can be tracked. This property has been exploited to estimate motion in a wide range of applications including in echocardiography and elasticity imaging [7], [10]. The current paper extents this approach to tongue imaging, thus providing point correspondences and additional robustness.

Traditional approaches to speckle tracking include block matching and optical flow-based schemes. Both schemes work well for sequences with small motion and simple deformations (e.g. rotation, translation, shear). Given the low frame-rate of commercial ultrasound scanners and the complex

motion of the tongue, they may fail to capture the temporal variations of the tongue. Moreover, these methods, being local in nature, treat each neighborhood independently. They are also not very robust to noise and artifacts. Hence, we adapt the registration based motion estimation framework, introduced in [1] to estimate the tongue motion. In contrast to traditional algorithms, this approach can model complex deformations and perform global estimation. It uses the information from the entire image to derive the fit. In contrast to [1], where each frame is deformed with respect to the first one, we deform each frame to the next one. This approach makes our approach more robust to speckle decorrelation, which may occur when scatterers move away from each other due to the large deformation of the tongue, and thus change the speckle pattern. Out of plane motion also changes the pattern and this change is also often referred as decorrelation.

The registration scheme estimates the deformation between consecutive frames; the combined deformation map enables us to track any point on a specified image in the sequence to other images. Specifically, one can track virtual markers on the tongue surface in any specific frame. Moreover, this approach can also derive the tongue contour information similar to the traditional approaches described in [4] and [6].

In order to illustrate our speckle tracking method, we tested it on ultrasound data acquired during speech. We obtained sequences of tongue curves during which the tongue transitioned from a velar stop consonant [g] in a preceding word to a word initial mid-back vowel [ɔ] in two different prosodic environments (word initial (Wd) and initial in an accentual phrase (AP)). We have chosen this test material as it shows the well-known phenomenon of domain initial strengthening [2]. The results confirm our expectations - based on previous research - that we find more displacement of the tongue in the higher prosodic domain (AP) compared to the word initial domain which is lower on the prosodic hierarchy.

2 Model based speckle tracking

The classical approach to track speckles to quantify motion information is block matching [3]. The main drawbacks of these local approaches are (a) constant regions can provide insufficient velocity cues, leading to poor estimation results (b) inability to introduce smoothness priors in the tracking process (c) local nature of tracking, leading to poor robustness. Hence, we adapt the non-rigid registration scheme, originally due to Carbayo et al. to quantify the motion information from ultrasound sequences [1].

2.1 Deformation model

We model the deformation between two consecutive images by the vector function:

$$\mathbf{g}(x,y) = \left(g_x(x,y), g_y(x,y)\right).$$

The functions g_x and g_y provide the co-ordinates in the target image, which correspond to x and y in the source image. Thus, the source image is deformed (by the deformation map $\mathbf{g}(x,y)$) to the target image:

$$I_s\left(g_x(x,y), g_y(x,y)\right) \approx I_t(x,y).$$

In contrast to [1], where they deformed every image to a the first image in the sequence, we map every image to the one immediately preceeding it:

$$I_{n-1}\left(\mathbf{g}_n(\mathbf{x})\right) \approx I_n(\mathbf{x}).$$

As discussed previously, this makes the algorithm more robust to speckle decorrelation. Once $\mathbf{g}_i(x,y), i = 0..n$ are obtained, the deformation of the tongue from the first image to the last is obtained as

$$\mathbf{g}_{n,1} = \mathbf{g}_n \circ \mathbf{g}_{n-1} \circ \ldots \mathbf{g}_1.$$

We model each of the deformations as Bspline functions, in terms of the coefficients $\{\mathbf{c},\mathbf{d}\}$, where

$$g_x(x,y) = x + \sum_{k,l=0}^{N-1} c_{k,l}\, \beta\left(\frac{x}{T} - k\right)\beta\left(\frac{y}{T} - l\right)$$

$$g_y(x,y) = y + \sum_{k,l=0}^{N-1} d_{k,l}\, \beta\left(\frac{x}{T} - k\right)\beta\left(\frac{y}{T} - l\right).$$

Here, β are cubic Bspline functions and T is the grid spacing. The main advantage of this representation is that the value and the derivatives of a Bspline model can be exactly evaluated and it satisfies multiresolution properties. Moreover, cubic Bsplines possess good approximation and smoothness properties.

2.2 Elastic registration for speckle tracking

The goal of the elastic registration scheme is to determine the deformation map between the source and target images I_{n-1} and I_n. We will use the sum of square differences between the deformed source image and the target image:

$$C_n = \left\| I_{n-1}\left(\mathbf{g}_n(\mathbf{x})\right) - I_n(\mathbf{x}) \right\|^2$$

as the criterion. The deformation is estimated as

$$\mathbf{g}_n = \arg\min_{\mathbf{g}} C_n .$$

Thanks to the Bspline deformation model, this boils down to the determination of the coefficients $\{\mathbf{c}_n, \mathbf{d}_n\}$. Since this is a non-linear optimization algorithm, we will use the steepest descend optimization algorithm to determine the optimal parameters.

3 Applying speckle tracking to linguistic data

3.1 Data acquisition and speech material

We applied our method to ultrasound images of the tongue. The images were acquired on a Siemens Anatres Sonoline ultrasound machine while the subject read the target sentences. The subject was a female native speaker of American English and her head was stabilized during data collection. Table 1 shows the sentences produced by the subject.

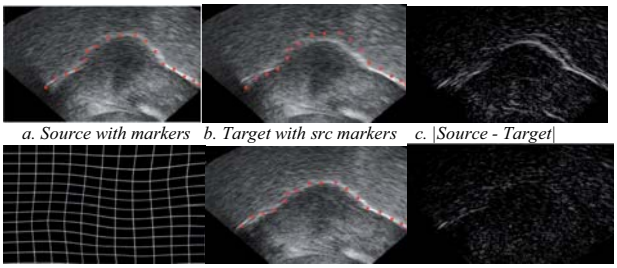
We used our method to compare velocity and displacement of the tongue during vowel production in the two different environments. Based on research describing domain initial strengthening [2], [5], we predict that the tongue is displaced more with respect to the preceding velar consonant in the prosodic domain that is higher on the prosodic hierarchy. Therefore, we expect more displacement in the AP condition compared to the Wd condition.

Table 1. *Test sentences in the two prosodic conditions with the target vowel underlined.*

	Test Sentences
AP	Silk, auk, and bolus are rare words.
Wd	The silk auk won't survive the winter.

3.2 Results

We will first illustrate the utility of the registration algorithm in tracking flesh point markers. We considered the first two frames of the Wd sequence in Fig. 1; these are the most challenging cases due to the large velocity of the tongue (see Fig 3). Virtual flesh point markers are added on the source image (1-a). The location of these markers in the target image (second frame) is shown in (b). Note that the markers are not aligned with the tongue surface. Using the registration algorithm, we estimated the deformation and tracked the location of the markers on the target image. Note from (f) that the deformed markers are in perfect alignment with the tongue surface in the target image. The low value of errors (absolute difference of deformed source image and target image) in (g) in comparison to (c) also indicates the accuracy of the registration. Note that the errors are at the noise level and no structures are visible.



a. Source with markers *b. Target with src markers* *c. |Source - Target|*

e. Deformation *f. Target with def. markers* *g. |Def. Src - Target|*

Figure 1: *Illustration of the registration algorithm. (a) is the source image with virtual tongue markers (b) is the target image (next frame) with the markers at the same location as the source. (c) is the absolute difference between the source and the target images. (e) is the estimated deformation and (f) is the target image with the deformed markers. (g) is the difference between the deformed source and the target images.*

Figure 2 shows the utility of the algorithm in estimating the horizontal and vertical displacement of the tongue during the AP and Wd sequences (from the consonant to the vowel). The AP sequence has 11 frames, while the Wd sequence has 4 frames. It is shown that the tongue settles for a lower vertical position at the vowel location in the AP sequence, in comparison to the Wd sequence. The top row is the horizontal and vertical displacements corresponding to the AP case. The bottom row indicates the displacements in the Wd case. Note that the displacements are larger in the AP case as expected.

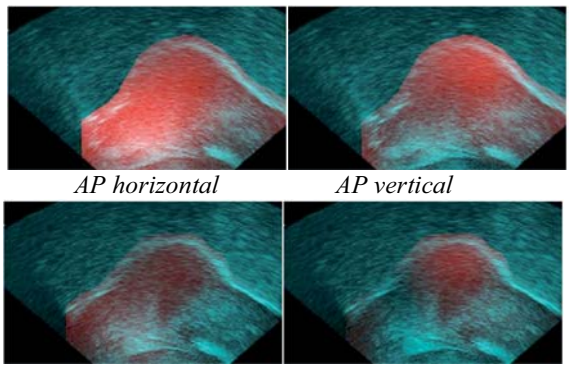


Figure 2: Comparison between the horizontal and vertical displacements of the tongue between the consonant and the vowel, estimated by the registration algorithm. The results are overlaid on the initial ultrasound image (during the consonant). Here, high intensity of red indicates large displacement values.

Figure 3 illustrates the estimation of the horizontal and vertical velocities of the tongue during the AP and Wd sequences. The top row is the horizontal and vertical displacements corresponding to the AP case, while the bottom row indicates the Wd case. Note that the estimated velocities are much larger in Wd case.

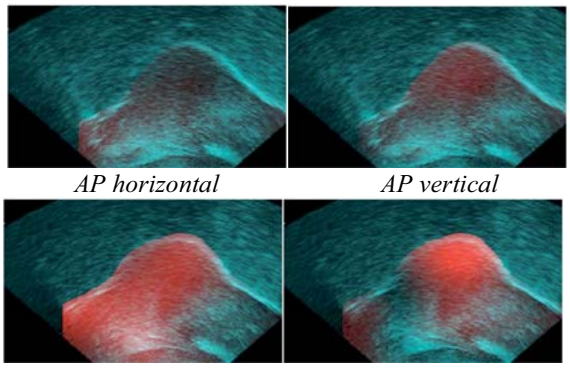


Figure 3: Comparison between the horizontal and vertical velocities of the tongue at the consonant location, estimated by the registration algorithm. The high intensity of red indicates large velocities.

4 Discussion and Conclusion

We have explored the use of speckle tracking/registration as a means of quantifying tongue motion from ultrasound image data. In contrast to standard approaches that track the tongue surface [4], we obtain a dense motion field. This enables us to track virtual flesh markers on the tongue to the subsequent images, enabling accurate quantitative comparison. We illustrated the utility of the algorithm in comparing two prosodic conditions. The estimated velocities and displacements are in agreement with expected results. In short, we found the proposed scheme to be a powerful tool in quantifying and analyzing tongue motion.

References

[1] M. Carbayo, et al. Spatio-temporal non-ridgid registration for ultrasound cardiac motion estimation. IEEE Tran. Med. Imag., 2005.

[2] C. Fougeron, & P.A. Keating. Articulatory strengthening at edges of prosodic domains. *JASA*, 101 (6): 3728-3740, 1997.

[3] I.A. Hien & W.O. Brien. Current time-domain methods for assessing tissue motion by analysis from reflected ultrasound echoes. *IEEE Tran. Ultrasonics*, 1993.

[4] K. Iskarous. Detecting the edge of the tongue. *Clin. Ling. & Phonetics*, 19 (6-7): 555-565, 2005.

[5] P.A. Keating. Phonetic encoding of prosodic structure. In: J. Harrington & M. Tabain. Speech Production. New York: Psychology Press, 2006.

[6] M. Li, C. Kambhamettu, & M. Stone. Tongue motion averaging from contour sequences. *Clinical Linguistics and Phonetics,* 19 (6-7): 515-528, 2005.

[7] G.E. Mailloux, et al. Restoration of the velocity field of the heart from two-dimensional echocardiograms. *IEEE Trans. on Medical Image Proc*., 8-6, 143-153, 1989.

[8] V. Parthasarathy, M. Stone & J.L. Prince. Spatiotemporal visualization of the tongue surface using ultrasound and kriging (SURFACES). *Clinical Linguistics and Phonetics*, 19 (6-7): 529-544, 2005.

[9] M. Stone. A Guide to Analyzing Tongue Motion from Ultrasound Images. *Clin. Ling. and Phonetics*, (19) 6-7, pp. 455-502, 2005.

[10] M. Sühling, et al. Myocardial motion analysis from B-mode echocardiograms. *IEEE Trans. on Image Processing*, 14:4, 525-536, 2005.

[11] D. H. Whalen, et al. The Haskins Optically Corrected Ultrasound System (HOCUS). *Journal of Speech, Language, and Hearing Research*, 48, pp.543-553, 2005.