DIRECT, NEAR REAL TIME ANIMATION OF A 3D TONGUE MODEL USING NON-INVASIVE ULTRASOUND IMAGES

Shicheng Chen[1], Yifeng Zheng[1], Chengrui Wu[1], Guorui Sheng[1], Pierre Roussel[2], Bruce Denby[1]

[1]Tianjin University, Tianjin, China
[2]Institut Langevin, Paris, France
coder.chen.shi.cheng@gmail.com, yzhen075@uottawa.ca,wcrtju01@tju.edu.cn
shengguorui@outlook.com, pierre.roussel@espci.fr, denby@ieee.org

ABSTRACT

A new technique for representing speech articulation with an ultrasound-driven finite element model of the tongue is presented. By using a snake contour extraction algorithm with anatomically motivated constraints and a common coordinate system between the ultrasound and the tongue model, it is possible for the first time to obtain a realistic 3D simulation of the tongue directly from a non-invasive sensor (ultrasound), without mapping through any intermediate sensor modalities, and at near real-time frame rates.

Index Terms— speech production, ultrasound, silent speech interface, active contours, dynamic programming

1. INTRODUCTION

Detailed information on tongue shape during articulation is of great interest in speech production research; for the treatment of speech pathologies; as an aid in language learning and rehabilitation; as well as for non-acoustic speech recognition for Silent Speech Interfaces (SSI) [1]. As most of the tongue is normally not visible, obtaining such information often involves invasive sensors such as electropalatographic (EPG) palate inserts [2], electromagnetic articulography (EMA) sensors glued to the tongue [3], or magnetic resonance imaging (MRI) [4]. Non-invasive ultrasound (US) tongue images are also widely used, but can be difficult to interpret due to speckle noise, inhomogeneous echogenicity, and experimental artifacts [5]. Representing a speaker's tongue motion in real time both accurately and non-invasively remains a challenging goal.

In this article, we show that by applying "anatomical constraints" to a dynamic snake tongue contour extraction algorithm, and imposing a common coordinate system between the ultrasound probe and a 3D Finite Element Model (FEM) of the tongue, the model can be driven in a realistic way directly from the ultrasound images.

2. RELATION TO PRIOR WORK

A number of approaches to real time modeling of tongue movement during articulation have appeared in the literature in the past few years. In the pioneering work of Yang and collaborators [6], constraint nodes in a tongue FEM were animated directly using coordinates of EMA sensors – which unfortunately must be glued onto the tongue. One non-invasive approach is to use the acoustic speech signal to estimate and display a speaker's most probable tongue movements, either by creating synthetic EMA coordinates to drive constraint nodes in a tongue model, as in [7], or by direct inversion of the speech signal, as in the talking head display of [8]. However, for applications in speech pathology and rehabilitation, language learning, SSI, and the like, deducing tongue movement from acoustic input is a serious difficulty. Too, to be most useful for these applications, a system should display a speaker's *actual* tongue movements, not simply the most probable ones.

Non-invasive ultrasound imaging may be a solution to such concerns. In [9], for example, a Gaussian Mixture Model (GMM) was used to map the first 20 principal components of ultrasound tongue images onto virtual EMA coordinates to drive a talking head. Clearly, though, it would be preferable to use ultrasound to animate an FEM directly, without passing through any intermediate modalities – basically replacing the EMA of [6] with ultrasound. A system to do this was proposed in [10][11], where unfortunately pathological configurations of the model were often encountered. In [12], such pathological behavior was avoided by matching an ultrasound contour to tongue configurations contained in a previously defined "dictionary", which unfortunately restricts the range of possible shapes, and in any event requires 1.2 seconds per frame to perform.

The method proposed here uses ultrasound images to directly animate a tongue FEM at near real time rates, by using anatomical considerations and ultrasound/model coordinate matching to control model instability. The tongue model and driving method are introduced in the following section, while the innovative aspects of this work are

presented in section 4. Results, including links to video demonstrations, appear in section 5, and some conclusions and perspectives for future improvements are given in the section 6.

3. TONGUE MODEL

3.1. Parameters and registration

The tongue FEM used has been described elsewhere, [6] [10][11][12]. It makes use of the Artisynth generic tongue mesh [13], further subdivided into 13,000 nodes and 44,000 tetrahedral elements. The at-rest sagittal contour of the model was aligned and scaled by hand to coincide as well as possible with a selected rest position ultrasound contour, including the hyoid position, placing the base of the tongue at the ultrasound probe.

3.2. Animation approach

To avoid the complexities of a muscle-driven model, animation was performed by assigning displacements to four constraint nodes, approximately equally spaced on the upper part of the mid-sagittal contour, avoiding the tongue tip, which does not image well in ultrasound. A fifth constraint point was chosen at the model hyoid, whose position in the mesh is known. The nodes are indicated in figure 1a by yellow dots on the sagittal contour, shown in pink, superimposed on an ultrasound tongue image. The base of the model tongue was secured by defining all nodes located there as anchor nodes. One iteration of the model animation consists of updating the coordinates of the constraint nodes with information obtained from the most recent ultrasound image, and solving the resulting new equations of motion. The update procedure is described in section 4.

3.1. Model stability

When driving a tongue FEM with a small number of constraint nodes, pathological configurations may arise in the following situations:

1. Two constraint points drift too close together, due to noise or contour tracking issues, causing "bulging" in the model due to its (approximate) incompressibility. Constraint nodes drifting too far apart, in turn, can lead to "drooping" of the model.
2. A too large constraint node excursion, arising from unstable contour extraction, may drive the model into a configuration from which it never recovers.
3. Constraint node displacement errors due to a mismatch of coordinate systems between the ultrasound image and the model can accumulate and lead to instability.
4. The model must be able to attain the constraint node displacement targets in the time allocated, to prevent accumulation of errors.

4. CONTOUR EXTRACTION AND MODEL UPDATE

The ultrasound contour extraction algorithm and model update procedure were designed to address the stability concerns mentioned above.

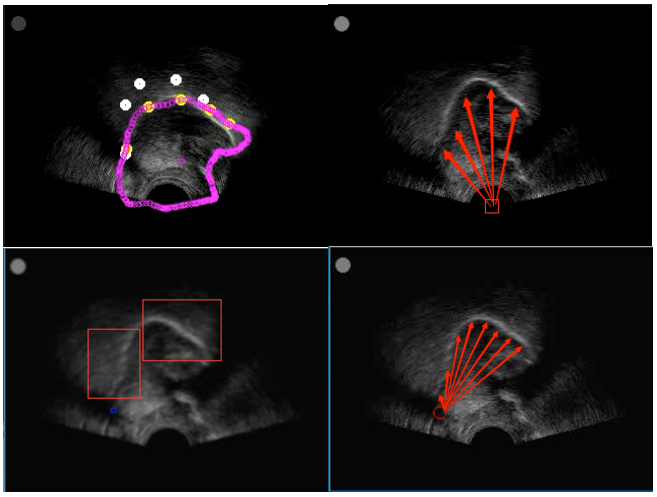

Figure 1. a) Pink curve and yellow dots are mid-sagittal contour and constraint nodes, respectively, of the tongue FEM. Snake points (upper white dots) are initialized in the upper part of image and attach themselves to the true contour in only a few frames; b) Distances from snake points to center of ultrasound transducer; c) Positive contour slope on left, negative on right; d) Ascending order of lines from snake points to hyoid bone.

4.1. Snake algorithm with "anatomical constraints"

The ultrasound data consist of disk files of tongue images acquired at 70 frames per second with a 4-8 MHz, 128 element microconvex probe maintained beneath the speaker's chin on an acquisition helmet. The tongue contour is represented by a 15-point snake [14] superimposed on the image, updated on each frame read from the file to reflect the tongue's movement since the last frame (the snake initialization procedure will be discussed below). To perform an update, the algorithm uses dynamic programming to select the 15 new snake points that both maximize the snake energy and satisfy a set of additional constraints. Snake energy consists of an external part, E_{ext}, related to the image intensity, and an internal part E_{int} determined by the internal configuration of the snake.

To calculate E_{ext}, the update algorithm considers new snake point candidates within 3 pixels of their previous positions. Each candidate makes a contribution to E_{ext} of $R + d$, where R is the total intensity in a 3 by 3 pixel block centered on the new point, and d the absolute distance in

pixels from the point to the center of the ultrasound transducer (see figure 1b). Maximizing *R* thus favors bright regions on the tongue contour that are consistent with typical tongue velocities at 70 frames per second; while maximizing *d* penalizes unphysical, spurious contours that might appear in the noisy lower part of the image.

Table I. Snake algorithm pseudocode

for $v_{i-2}v_{i-1}v_i$ (i:=2-14) **begin**
for v_i in search region **begin**
$R_{u,v} = \sum_{i=-1}^{1}\sum_{j=-1}^{1} I_{u+i,v+j}$ (intensity 3x3 pix. block)
for v_{i-1} in search area **begin**
for v_{i-2} in search area **begin**
if (points too close or far) **continue**
if (slope not pos. LHS) **continue**
if (slope not neg. RHS) **continue**
if (hyoid slopes not ascending order) **continue**
$E_{ext}(v_i) = \mid v_{center} - v_i \mid$
$E_{\text{int}}(v_i v_{i-1} v_{i-2}) = \arctan(\frac{vy_{i+1} - vy_i}{vx_{i+1} - vx_i}) - \arctan(\frac{vy_i - vy_{i-1}}{vx_i - vx_{i-1}})$
{update dyn. prog. array $dp[v_i][\text{pos}\,v_i][\text{pos}\,v_{i-1}]$
by E_{ext} and E_{int} if > orig. val.} **end end end end**

E_{int} is given by the sum of the angles between adjacent snake segments, thus favoring "smooth" contours. The additional constraints below, related to the internal configuration of the snake but not included in E_{int}, are easily handled with dynamic programming, see the algorithm pseudocode in Table I.

- Distances between adjacent snake points must be between 7 and 11 pixels inclusive, to inhibit unnatural compression or stretching of the tongue;
- A positive slope is required on the left of the contour, and a negative one on the right (see figure 1c), corresponding roughly to the shapes of the tongue root and dorsum regions;
- Slopes of lines drawn from snake points to the hyoid bone must have ascending order, see figure 1d. This constraint both penalizes unphysical tongue compression and allows for rapid algorithm recovery after swallowing, which otherwise causes the snake to break loose from the contour. The hyoid location is easily determined by finding the darkest 5 x 5 pixel area in the hyoid shadow region of the ultrasound image.

These "anatomical" constraints, which can be considered as deriving both from the properties of the actual speaker's tongue in the ultrasound image, and the physical model that is supposed to represent it, are quite important for the quality of the contour tracking.

To initialize the snake, 15 points forming a suitably shaped curve are placed in the upper part of the first image, away from the tongue body. The snake automatically attaches itself to the true tongue contour within a few frames during the normal running of the algorithm, see figure 1a, where four of these initial snake points appear as white dots.

A comparison of performance to the popular EdgeTrak [5] algorithm, which also uses a snake approach, has not been made; however, the manual initialization and periodic resets necessary with EdgeTrak would make its use for driving a model rather more difficult.

4.2. Model update procedure

Once the snake points have been updated, a new set of constraint node coordinates must be furnished to the tongue model so that it may evolve to its new configuration. Snake points numbered 3, 6, 10, and 14 (proceeding left to right in the image), as well as the hyoid point, were chosen to create input for the model; these appear as white dots in figure 1a at their initialized locations. At each new image, the snake and hyoid locator algorithms provide the horizontal and vertical displacement values of these 5 points, dx_i and dy_i, i = 0-4 (with 0 the hyoid point), relative to the preceding frame, and expressed in pixels on the ultrasound image.

The corresponding displacements of the model constraint nodes are defined as $\Delta x_i = \alpha dx_i$ and $\Delta y_i = \alpha dy_i$, where α is a fixed scale factor, and the constraint nodes are also numbered 0-4 left to right. Although snake points, contrary to the constraint nodes, are not tissue points, this simple procedure nonetheless gives reasonable results. The value of α was determined manually (final value α = 0.015), so as to make the movement of the model sagittal contour coincide as closely as possible with the movement of the ultrasound contour and hyoid bone.

4.3. Implementation

Due to the rather different computing environments of the model and snake programs, the two algorithms presently run on separate computers, communicating over a *tcp* socket. In the current setup, the system is implemented in the following way:

1) An ultrasound image is read in from the file by the contour tracking program, running on a Dell workstation.
2) The snake algorithm runs until completion.
3) The resulting dx_i and dy_i values are sent to a socket via tcp.
4) The tongue model computer, a Mac Pro, retrieves dx_i and dy_i, calculates Δx_i and Δy_i, and simulates until the model update is completed.
5) Repeat from 1).

Due to overheads from model (60%) and snake (40%) processing, the current free-running throughput of the system, implemented in the way described above, is 21

frames per second, that is, only about a factor of 3 away from the original 70 Hz acquisition rate.

5. RESULTS AND DISCUSSION

A short video of the results, showing the model and the ultrasound frame with snake and model sagittal contour superimposed, can be accessed at [15]. The speaker pronounces a few isolated /g/ and /t/ and then repeats the word "escalator" several times. The video has been speeded up to the original 70 frames per second ultrasound acquisition rate in order to accurately represent the tongue motion. It is seen that the model follows the movement of the snake rather faithfully – apart from some obvious artifacts – even for large excursions of the tongue. Four stills from the video appear in Figure 2. A second, longer version of the 70 Hz video, showing just ultrasound and sagittal contour, can be viewed at [16], where the pink horizontal and vertical lines in the images are estimates of the palate and upper teeth positions, respectively.

While the performance of the system can certainly be improved, we believe this to be the first time a 3D tongue model has been directly driven in a believable way at near real time using a non-invasive sensor – in this case ultrasound.

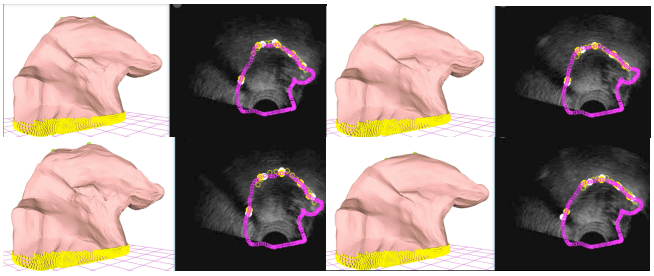

Figure 2. Four screen captures of the tongue model driven by ultrasound images during speech.

6. CONCLUSIONS AND PERSPECTIVES

By using a snake algorithm with anatomically motivated constraints, and a common coordinate system between ultrasound images and a 3D tongue model, we have shown that realistic, direct animation of a tongue model from a non-invasive source is possible by assigning snake point movements to a small set of constraint nodes. The system currently achieves a throughput of 21 frames per second. If the model and ultrasound acquisition can be merged onto a single, more powerful computational platform, it should be possible to drive the model directly from 70 Hz ultrasound. The snake tracking algorithm works well on speakers with tongues not too dissimilar to the default model. Elastic registration of the model mesh to individual speakers will allow the tests to be properly extended to additional speakers in the future. More elaborate animation schemes that better approximate a muscle-driven organ are also to be investigated.

7. ACKNOWLEDGMENT

The authors thank Professor Yin Yang of the University of New Mexico, U.S.A., for providing access to the 3D model code as well as for advice on its utilization. Partial funding for this work was provided by the China Ministry of Education "985 Foundation" via grant number 060-0903071001. Thanks go also to the reviewers, who provided much valuable input for improving the article.

8. REFERENCES

[1] B. Denby, T. Schultz, K. Honda, T. Hueber, J.M. Gilbert, J.S. Brumberg, "Silent Speech Interfaces", *Speech Communication*, vol. 52, pp. 270-287, 2010.

[2] B. Bernhardt, B. Gick, P. Bacsfalvi, J. Ashdown, "Speech habilitation of hard of hearing adolescents using electropalatography and ultrasound as evaluated by trained listeners", *Clinical Linguistics and Phonetics*, 17, 199-216, 2003.

[3] M. Kuruvilla, B. Murdoch, J. Goozèe, "Electromagnetic articulography assessment of articulatory function in adults with dysarthria following traumatic brain injury", *Brain Injury*, 21(6):601-13, 2007.

[4] V. Parthasarathy, J.L. Prince, M. Stone, E.Z. Murano, M. Nessaiver, "Measuring tongue motion from tagged cine-MRI using harmonic phase (HARP) processing", *Journal of the Acoustical Society of America*, vol. 121, issue 1, pp. 491-504, 2007.

[5] Y.S. Akgul, C. Kambhamettu, M. Stone, "Automatic extraction and tracking of the tongue contours", *IEEE Transactions on Medical Imaging*, 18, 1035-1045, 1999.

[6] Y. Yang, X. Guo, J. Vick, L.G. Torres, T.F. Campbell, "Physics based deformable tongue visualization", *IEEE Transactions on Computer Graphics*, 19(5) 811-823, 2013.

[7] R. Luo, Q. Fang, J. Wei, W. Lu, W. Xu, Y. Yang, "Acoustic VR in the mouth: A real-time speech-driven visual tongue system", *IEEE Virtual Reality Conference*, Los Angeles, USA, 2017, pp. 112-121.

[8] T. Hueber, G. Bailly, P. Badin, P. Elisei, "Speaker adaptation of an acoustic-articulatory inversion model using cascaded Gaussian mixture regressions", in *Proc. of Interspeech*, Lyon, France, 2013, pp. 2753-2757.

[9] D. Fabre, T. Hueber, P. Badin, "Automatic animation of an articulatory tongue model from ultrasound images using Gaussian mixture regression", *Proc. of Interspeech*, Singapore, 2014, pp. 2293-2297.

[10] K. Xu, Y. Yang, A. Jaumard-Hakoun, M. Adda-Decker, A. Amelot, S.K. Al Kork, L. Crevier-Buchman, P. Chawah, G. Dreyfus, T. Fux, C. Pillot-Loiseau, P. Roussel, M. Stone, B. Denby, "3D tongue motion visualization based on ultrasound tongue image sequences", in *Proc. of Interspeech*, Singapore, 2014.

[11] K. Xu, Y. Yang, A. Jaumard-Hakoun, C. Leboullenger, G. Dreyfus, P. Roussel, M. Stone, B. Denby, "Development of a 3D tongue motion visualization platform based on ultrasound image sequences, *Proc. of ICPhS*, Edinburgh, UK, 2015.

[12] K. Xu, Y. Yang, C. Leboullenger, P. Roussel, B. Denby, “Contour based 3D tongue motion visualization using ultrasound image sequences”, *Proceedings of ICASSP*, Shanghai, China, 2016.
[13] J.E. Lloyd, I. Stavness, S. Fels, “ArtiSynth: A fast interactive biomechanical modeling toolkit combining multibody and finite element simulation”, *Soft tissue biomechanical modeling for computer assisted surgery*, Springer, 355-394, 2012.
[14] M. Kass, A. Witkin, D. Terzopoulos, “Snakes: Active contour models”, *International Journal of Computer Vision*, 1.4, 321-331, 1988.
[15] https://youtu.be/FCq2akoB_O8
[16] https://youtu.be/L-6gW1UZS6Q