MULTIMODALITY ACQUISITION OF ARTICULATORY DATA AND PROCESSING

Michael Aron[1], Anastasios Roussos[2], Marie-Odile Berger[1], Erwan Kerrien[1], Petros Maragos[2]

[1] LORIA/ INRIA Grand Est, BP 101, 54602 Villers les Nancy, France
email: {aron,berger,kerrien}@loria.fr
[2] School of ECE, National Technical University of Athens, Greece
email : {troussos,maragos}@cs.ntua.gr

ABSTRACT

In this paper[1], a framework to acquire and process dynamic data of the tongue during speech processing is presented. First, a setup to acquire data of the tongue shape combining ultrasound images, electromagnetic localization sensors and sound is presented. Techniques to automatically calibrate and synchronize the data are described. A method to extract the tongue shape is then proposed, by combining a preprocessing of the ultrasound images with an image-based tracking method that integrates adapted constraints.

1. INTRODUCTION

Being able to build a model of a speaker's vocal tract and face is a major breakthrough in speech research and technology. A vocal tract representation of a speech signal would be both beneficial from a theoretical point of view and practically useful in many speech processing applications (language learning, automatic speech processing, speech coding, speech therapy, film industry...).. This requires not only to design an acquisition system but also to define appropriate image processing techniques to extract the articulators (tongue, palate, lips...) from the data.

An ideal imaging system should cover the whole vocal tract (from larynx to lips) and the face, have a sufficient spatial and time resolution, and not involve any health hazard. At present, no single imaging technique answers the above requirements alone: the dynamics of the tongue can be acquired through ultrasound (US) imaging [12, 10] with a high frame rate but these 2D images are very noisy; 3D images of all articulators can be obtained with magnetic resonance imaging (MRI) but only for sustained sounds [5], electromagnetic (EM) sensors enable the investigation of speech articulators dynamics for a small number of points of the tongue [11]. Therefore, combining several imaging techniques is necessary. Several attempts have been made to acquire multimodal articulatory data. The HOCUS system [15] combines US imaging together with infrared emitting diodes placed on the lips and on the probe. In the HATS system [12], M. Stone used several 2D US acquisitions to recover a 3D model of the tongue. Using multimodality requires to perform spatial calibration and temporal synchronization in order to express the data in the same spatio-temporal frame. Though very important, these aspects are generally either not addressed in the above mentioned systems or at best manually performed.

The system we foresee will integrate 3D MRI, high-speed stereo-vision, US imaging and EM sensors to produce 3D+t, i.e. dynamic, models of the vocal tract. We focus here on the dynamic part, that US imaging and EM sensor form, to get the tongue shape, including the apex.

Our contributions in this paper are twofold. First, section 2 gives an overview of our US+EM acquisition system. In addition, fully automatic procedures for spatial calibration and synchronization are described. Our second contribution is on the processing of the articulatory data. We here focused on the processing of US images which bring important information on the tongue dynamics. Despite pioneering efforts conducted by M. Stone [12, 10], there does not exist any efficient tool to track the tongue shape over US sequences with sufficient reliability. We thus describe in section 3 efficient pre-processing methods of US images to reduce noise and to enhance the tongue contours. A complete framework to automatically extract the tongue shape over sequences which takes advantage of the EM sensors is then described. Significant results are given in section 4.

[1]The authors acknowledge the financial support of the Future and Emerging Technologies (FET) programme within the Sixth Framework Programme for Research of the European Commission, under FET-Open contract no. 021324".

2. THE ACQUISITION SYSTEM

2.1 General setup

US imaging provides a continuous tongue shape in the mid-sagittal plane. However, the air in the sublingual cavity and/or the jaw bone block the US signal, making the apex invisible in most of the US images. One EM sensor glued on the apex of the tongue thus provides point-wise tracking of the extremity of the tongue. A second EM sensor glued on the tongue dorsum helps to constrain the location of the tongue contour in US images.

Our setup relies on a Logiq5 Expert US machine (GE Healthcare) and the Aurora EM system (NDI). The speaker's voice is recorded by a microphone connected to a PC. The main characteristics of each modality are summarized in table 2.1. In cine loop mode, the acquired US images are stored in a video buffer. This buffer is saved to disk in DICOM format when the user presses a footswitch at the end of the sequence. The US machine was tuned to acquire 532×434 pixels images (approx 9 cm $\times$ 7 cm) at 66 Hz, which correspond to a 15 seconds video buffer.

	EM	US	Sound
Acquisition rate	40 Hz	66 Hz	44100 Hz
Recording time	unlimited	15 seconds	unlimited
File data format	text	DICOM	WAV
Recording process	real time	cine loop	real time

Table 1: Main characteristics of the modalities used in our acquisition system.

2.2 Synchronization

US images, EM sensor data and sound must be synchronized to enable data fusion. Fig. 1 depicts the complete acquisition system together with synchronization add-ons. The system core is a control PC which provides a precise time-line to stamp the various signals: it sends a stop signal to the US machine to save the video buffer to disk, thanks to an electronic relay that simulates a footswitch pressure; it sends timestamped start and stop signals to the Aurora EM system; and it emits beeps at the start and end of the acquisition to timestamp the sound data.

Sending start and stop signals allows us to measure the actual acquisition frequency and drastically reduce the temporal shift. For example, 2% error is common on sound acquisition rate (e.g. 43.2kHz instead of 44.1 kHz). Such an error implies a 300 ms shift at the end of a 15 seconds acquisition, that is a shift by 20 US images in our setup.

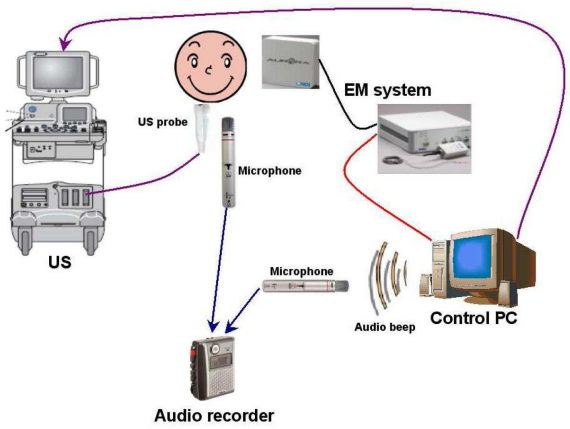


Figure 1: US+EM system setup.

Experiments were conducted to assess the synchronization accuracy of our setup. The maximum time lapse is below 1 ms for sound and 25 ms for EM data. Since we had no access granted to the internals of the US machine, we could only measure the repeatability of the time delay between the stop signal and the last US frame actually captured. The average delay was measured to be 52 ms, with a standard deviation of 14 ms (approx. 2 US images). Further details can be found in [1].

2.3 US/EM spatial calibration

The system comprises 2 spatial modalities: EM data and US images. US images are dependent on the probe position, and therefore may move along the acquisition. On the other hand, EM sensors are located relative to an EM field generator whose position is fixed. Spatial data fusion thus requires that the US images are expressed in the fixed EM reference frame. Therefore, the US probe position is tracked thanks to a 6DOF EM sensor connected to the Aurora system. The unknown, but fixed, linear relation between the US image inner reference frame and the 6DOF sensor local reference frame is calibrated by imaging a calibration fantom under various probe positions: a wooden stick whose line equation is known in the EM frame, and whose intersection with the US image plane is visible in the US images. Further details can be found in [2].

After calibration, tracking the 6DOF EM sensor becomes equivalent to tracking the US image plane. This latter plane can then be positioned with respect to the EM reference frame in real time during the acquisition. It is thereby possible to check in real time that the EM sensors glued on the tongue lie within the US image plane, which is necessary to fuse the acquired data. A warning for US/EM misalignment was integrated in our system.

Finally, the tongue motion, as seen in US images, may be corrupted by global head motion. Indeed, in our experience, too strong head immobilization hampers natural speech utterance over long acquisition sessions. Therefore, our system does not include any head immobilization devices. Instead, the speaker is asked to remain still. The head position is corrected for occasional motion thanks to 2 5DOF EM sensors glued behind the speaker's ears.

3. TRACKING THE TONGUE SHAPE IN THE US SEQUENCES

Our goal is to develop automatic tools (or as automatic as possible) in order to extract the tongue shape in the vast amount of US sequences recorded in our corpus. Many works have been devoted to tracking in the computer vision community. However tracking in US sequences is well known to be difficult due to the high speckle noise. There are many other difficulties specific to our application: (i) the high speed of the tongue for some sounds produces blurred contours which are difficult to track, (ii) the tongue undergoes high elastic deformations (ii) the apex is often not visible as the air of the sub-lingual cavity prevents sound propagation. To our knowledge, the specific problem of tongue tracking in US images has only been addressed by M. Stone et al. who developed the EdgeTrak software [9]. Tracking in EdgeTrak is based on a snake model which incorporates edge gradient and intensity information as well as information on contour orientation. Thorough experiments with this software were conducted on our database and showed significant drawbacks: tracking failures often occurred and needed manual correction which prevented us from using this method to process long sequences of US data. These failures are mainly due to the fact that the dynamics of the tongue is not exploited in the tracking scheme.

To cope with these problems, we have extended one of our previous works on contour tracking in US images [3] to the particular case of the tongue. As in [4, 7], the general idea of the tracking process was to restrict the class of allowable motions (shape deformations) to an admissible set of parametric motions in order to reduce the influence of noise.

The contour is manually initialized in the first frame of the sequence. Once initialized, the tracker operates in a loop following two major stages: a prediction of the contour is computed under the hypothesis of a parametric model. A refinement stage based on snakes [8] is then performed to compute the actual position. Adaptation of this general scheme to cope with the specificities of the tongue and to improve the robustness of the tracking includes the following:

- An original preprocessing method that reveals the contour of the tongue and is robust against speckle noise.
- Prediction of the position is performed under the hypothesis of a parametric motion model (affine model) using the optic flow to meet the required robustness.
- Appropriate boundary conditions for the snake in the refinement stage are applied.
- Robustness of the tracking is enforced by introducing prior on the tongue shape provided by the EM sensors glued on the tongue.

These improvements are described in more details in the rest of the section. In the following, US images are presented so that the apex is on the left and the palate on top.

3.1 Overview of the visual tracking

Robustness of the tracking procedure is achieved by restricting in a first step the class of allowable tongue deformation/motion to an admissible set. The prediction of the contour is then computed thanks to the optic flow between two consecutive images. Since only the normal component of the optic flow [6] is reliable, the motion f_p is computed as the one minimizing (see Fig.2)

$$\sum_{1\leq i<N} |(\overrightarrow{M_i f_p(M_i)}.n_i)n_i - v^{\perp}(M_i)|^2 \qquad (1)$$

Where $\{M_i\}_{1\leq i\leq N}$ are the points of the contour C to be tracked, p is the set of parameters (6 for an affine motion), n_i is the unit normal to C at point M_i and $v^{\perp}$ is the normal component of the optic flow. As the optical flow does not match the true displacement, we refine the estimation iteratively in the following way: we compute the normal optical flow $v_1^{\perp}$ on the curve $f_{\hat{p}_0}(M_i)_{\{1\leq i\leq N\}}$ between the registered image $I_1(f_{\hat{p}_0}^{-1})$ and I_2. Resolving (1) with $v = v_1$ then allows us to compute $f_{\hat{p}_1}$. Hence, $f_{\hat{p}_1} \circ f_{\hat{p}_0}$ is a better estimate of the motion and so on... Successive infinitesimal refinements $f_{\hat{p}_q} \circ ... \circ f_{\hat{p}_0}$ give an accurate estimation of the motion field. The prediction obtained is generally close to the actual contour.

Then, the constraint on the deformation is dropped and the position of the contour is refined by using the classical snake model [8]. For a parameterization s of a contour C, it consists in minimizing the energy functional:

$$\phi(C) = \int_0^1 E_{img}(C) + E_{int}(C) ds \qquad (2)$$

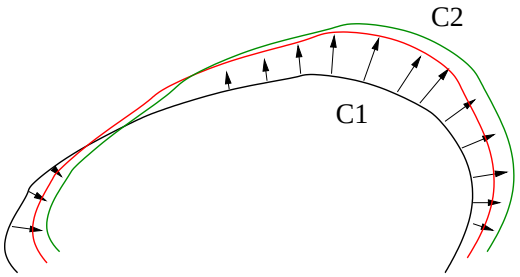


Figure 2: Predicting the tongue shape: given the contour C_1 in the current image (in black), the affine transformation that best matches the normal optic flow (black arrows) is iteratively computed giving rise to the predicted curve (in red). This curve is generally close to the searched contour C_2 (in green) and is likely to converge using snake process.

The internal energy term E_{int} controls the snake shape and the image energy term E_{img} defines the features in the image that are of interest. The derivation of the latter term is the scope of the next section.

3.2 Preprocessing of the ultrasound images

Ideally, the image energy term $E_{\text{img}}(x,y)$ used in (2) should have local minima at the points of the curve to be tracked, which in our case is the lower border of the tongue. The most common choice is to derive this term from the simple *gradient image* $|\nabla G_\sigma * u(x,y)|$ computed on a gaussian smoothed version [2] of the intensity image $u(x,y)$ [8]. As section 4.2.2 reveals, this choice may be not as effective as usual, because of the high amount of speckle noise in US images.

Therefore, we have designed a sophisticated processing of US images, which simplifies them, emphasizes the tongue contour and extracts an effective $E_{\text{img}}(x,y)$. This method exploits the fact that the tongue's visible part is the lower border of a band with relatively high intensity. It processes each frame $u(x,y)$ of the US sequence separately and consists of the following steps:

1. $u(x,y)$ is converted to $u(r,\phi)$, which is the image's representation using the polar coordinate system with origin the intersection point of the US beams. This representation seems more convenient because the r-direction is exactly the direction of the US beam.

2. A robust estimation of the orientation $\theta(r,\phi)$ that is perpendicular to the edges of $u(r,\phi)$ is computed at every point. This is done by first computing the *structure tensor* of $u(r,\phi)$ [14]:

$$\mathbf{J}(r,\phi) = G_\rho * \left(\nabla(G_\sigma * u) \otimes \nabla(G_\sigma * u)\right), \qquad (3)$$

where "$\otimes$" denotes tensor product. Then, $\theta(r,\phi) \in [0,\pi)$ is derived as the orientation of the eigenvector that corresponds to the largest eigenvalue of $\mathbf{J}(r,\phi)$. Due to the convolutions in (3), $\mathbf{J}$ is insensitive to image details smaller than $O(\sigma)$ and is affected by the image variation within a neighborhood of size $O(\rho)$. Thus, σ must be relatively small, comparable to the characteristic size of speckle pattern and ρ must be larger than σ, comparable to the size of the tongue's bright band in $u(r,\phi)$.

3. $u(r,\phi)$ is correlated with a spatially varying kernel $k(r,\phi;r',\phi')$ (see Fig. 3.b), which in each point is aligned to the direction of $\theta(r,\phi)$:

$$f_1(r,\phi) = \int\int u(r',\phi')k(r,\phi;r+r',\phi+\phi')\mathrm{d}r'\mathrm{d}\phi'. \qquad (4)$$

The response $f_1(r,\phi)$ has large values mainly at the lower borders of bright bands. More precisely, the kernel k is constructed as follows:

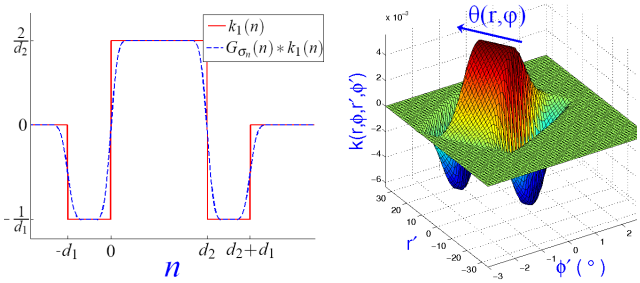
First, we consider the piecewise constant 1D kernel $k_1(n)$ that is plotted in Fig. 3. This kernel corresponds to the variation of an ideal bright band, in its normal direction n (the point $n=0$ corresponds to the lower border of this band). Therefore, d_2 should approximate the typical size of the tongue's bright band in $u(r,\phi)$. Also, we chose $d_1 = 0.44d_2$.

Afterwards, a regularized and 2D-extended version of $k_1(n)$ is constructed:

$$k_2(n,\xi) = \left(G_{\sigma_n}(n) * k_1(n)\right) G_{\sigma_\xi}(\xi)$$

where we chose $\sigma_n = d_1/8$ and $\sigma_\xi = d_1$ (see also Fig. 3.a). The convolution with G_{σ_n} smoothes the intra-region transitions. Also, the extension to ξ-direction makes the response f_1 more robust to speckle patterns. σ_ξ is relatively large, since near the tongue, the image variations in ξ-direction are caused only from the speckle. Roughly, the gaussians G_{σ_n} and G_{σ_ξ} offer an adaptive anisotropic smoothing, mainly aligned to the edges of $u(r,\phi)$.

Finally, the kernel $k(r,\phi;r',\phi')$ (Fig. 3.b) is derived from rotating the function $k_2(r',\phi')$ by the angle $\theta(r,\phi)$.



(a) (b)

Figure 3: Construction of the kernel $k(r,\phi;r',\phi')$, which is correlated with the US image frames.

4. We get $f_2(r,\phi)$ from $f_1(r,\phi)$ by setting to 0 all the negative values of f_1. This is done because, at the points with $f_1 < 0$, the image variation is closer to the pattern of a dark band, rather than a bright band. Thus, f_2 avoids the negative local maxima of f_1, which could undesirably attract the snake.

5. $f_2(r,\phi)$ is converted to $f_2(x,y)$. Then, $f_3(x,y)$ is computed as the *grayscale area opening* of $f_2(x,y)$ at a relatively small size scale A [13]. This operation "eliminates" the bright regions with area smaller than A. Usually, these regions are caused by speckle noise, since the tongue's bands have much bigger area.

6. Finally, the energy term is computed from $E_{\text{img}}(x,y) = -f_3(x,y)$, since it must have local minima, instead of maxima, at the points of the tongue contour.

As seen in Fig. 6, our method yields a result that reveals better the tongue and is less affected by speckle than the gradient image.

3.3 Boundary conditions on the snake

As it is well known, snakes naturally tend to shrink. Classical boundary conditions are fixed or free extremities which are not well suited to our problem. We have tested two different boundary conditions, defined by rays on Fig.4.a.

- the two snake extremities belong to the rays (in blue) defined by the extremities of the initialization curve.
- the left extremity belongs to the ray defined by the sensor position of the apex and the right extremity belongs to the ray defined by the right extremity of the initialisation curve.

Doing this, we avoid snake shrinking while letting the extremities free to be anywhere on the rays. Technically speaking, we use a polar coordinate system (r,ϕ), as in section 3.2, to represent the images, so that the rays are parallel. Our new boundary conditions can then be expressed as ϕ=*constant and r is free* which are boundary conditions easy to integrate in the snake process.

3.4 Tracking with US images and EM sensors

When images are very noisy, algorithms which are purely based on low-level properties will fail to detect the right contour. Incorpo-

[2] $G_\sigma(x_1,..,x_N)$ denotes an N-D isotropic gaussian kernel of standard deviation σ and "$*$" denotes convolution.

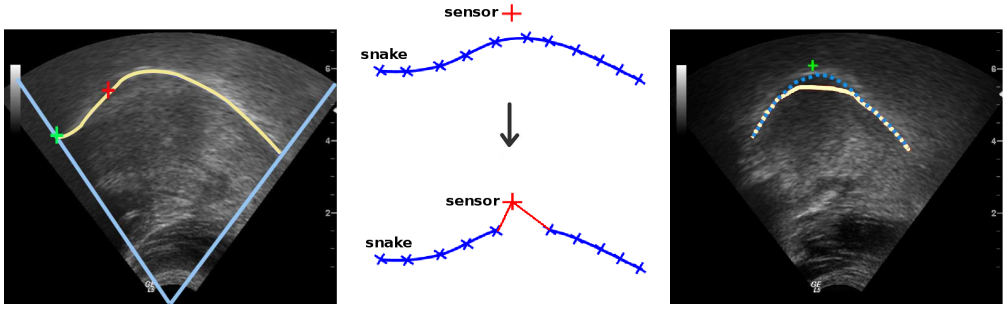


Figure 4: Improving the snake model: (a): definition of appropriate boundary conditions: the extremities belong to the rays drawn in blue (b) Reinitialization of the snake with the sensors (c) Comparison of sensorless (yellow) vs. sensor-added (blue) visual tracking (sensor = green cross).

rating shape prior is then a good mean to constrain the tracking algorithm to detect natural tongue shapes. In this work, the locations given by the sensors glued on the tongue are used as prior. Due to slight registration or synchronization errors, the sensors do not always belong to the surface of the tongue. We therefore prefer to use these positions as soft constraints instead of constraining the snake to pass through these points. A first solution is to create attractive force fields towards these positions. A simpler but more efficient solution is to modify the predicted curve by integrating the sensors in the predicted curve as shown in Fig.4.b. This way, the snake is locally attracted to the gradient created by the tongue and this often dramatically improves the convergence of the snake. The interest of such a strategy is demonstrated in Fig. 4.c.

4. RESULTS

4.1 Acquisition system

A corpus of US sequences with EM and audio data was successfully acquired with our acquisition system. This representative sample of speech data, during 10 min and 15 seconds, includes Vowel-Vowel (/ae/, /ai/, /yo/...), Vowel-Consonant-Vowel (/aka/, /isu/...) and complete sentences ("La poire est un fruit à pépins"...) in French.

4.2 Tracking method

4.2.1 Data

A sequence of 390 images (approximately 6 seconds) acquired by our system has been taken to evaluate our tracking method. This sequence includes four groups of phonemes (/ae/ /ai/ /ao/ /au/). On the first two groups of phonemes (/ae/ /ai/, 200 images), the tongue contour appears strong and the tongue motion is limited. On the last two groups (/ao/ /au, 190 images), large parts of the tongue contour are weak because the tongue is moving fast and is not correctly imaged by the US system.

In our experimentations, the affine motion model proved to be a good choice for the estimation of the displacement. In practice, the infinitesimal displacements rapidly converged towards identity. 50 iterations were used to estimate the global motion. Fig. 5 shows images with the tongue successfully tracked using our proposed algorithm. Whole video sequences which these images are extracted from are available on our website: http://magrit.loria.fr/Confs/Eusipco08/ .

4.2.2 Evaluation

Five tracking methods were evaluated :

- method 1: EdgeTrak
- method 2: snake on gradient images and extremities defined by the initialization
- method 3: snake on gradient images and use of the sensors
- method 4: snake on preprocessed images and extremities defined by the initialization

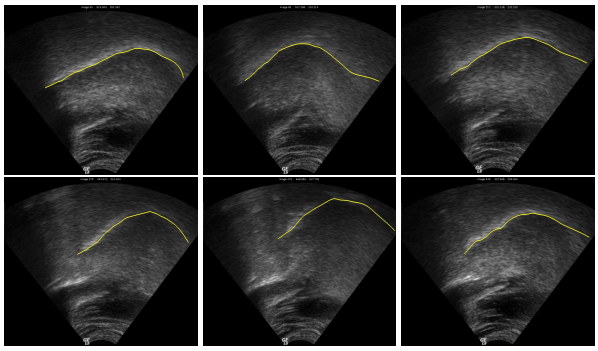


Figure 5: Tracking on 6 images of the sequence /ae/ /ai/ /ao/ /au/. Associated phonemes (left to right, top to bottom): /a/ in /ae/, /e/, /a/ in /ao/, /o/, /u/ (start), /u/ (end)

- method 5 : snake on preprocessed images and use of the sensors (our proposed method)

A manually drawn contour was considered as the ground truth position of the tongue shape to evaluate these tracking methods. The error on a curve was computed as the mean distance between each point on the curve and the closest point on the ground truth contour. The evaluation is based on the mean error (in mm) on whole sequences with the associated standard deviation. We also used the percentage of images of the sequences for which the error was above 2 mm. In our experimentations, this value appeared to be critical to consider the tracking failed. Results on the first two groups of phonemes are presented in table 2.

Method #	1	2	3	4	5
Mean error (mm)	1.36	1.18	1.34	1.45	1.31
Std. dev. (mm)	0.58	0.51	0.65	1.37	1.21
% images error $>$ 2 mm	17	9.5	14.5	14.5	14

Table 2: Results of the tracking on the group of phonemes /ae/ and /ai/ (200 images - 3 sec).

All methods yield similar mean errors for this sequence (between 1 mm and 1.5 mm): this sequence presents no major difficulties because the tongue contours are strong in the images. Fig. 6 shows the results of methods 3 and 5 for a frame of the sequence. We see that the tracking using our filtering to derive E_{img} instead of simple gradient yields an improvement to the shape of the tongue. But in both cases, the right part of the curve is not accurate because of the very low contrast of the tongue edge.

Our method has also been tested on the group of phonemes /ao/ /au/, where the tongue is moving faster.

Results presented in table 3 demonstrate the benefits of our method. The motion estimation brings robustness to the tracking

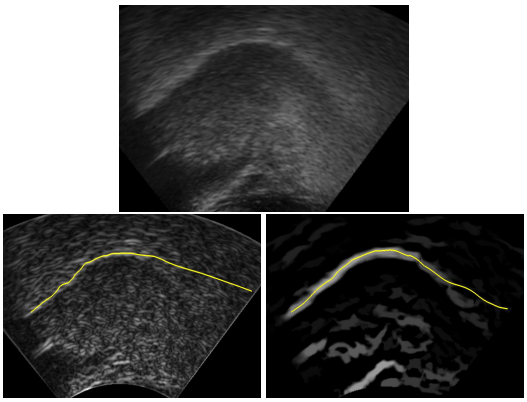


Figure 6: Image term E_{img} using the simple gradient image (left) and our preprocessing method (right). The corresponding tracking results of methods 3 (left) and 5 (right) are superimposed.

Method #	1	2	3	4	5
Mean error (mm)	5.68	1.83	1.79	1.72	0.97
Std. dev.(mm)	2.57	0.51	0.56	1.11	0.34
% images error $>$ 2 mm	93.2	35.3	34.2	18.9	1.6

Table 3: Results of the tracking on the group of phonemes /ao/ and /au/ (190 images - 2.9 sec).

with mean error below 2 mm. On the contrary, EdgeTrak fails in this difficult case because of the fast motion and the lack of constraints at the extremities. Sensors significantly improve the tracking, because the snake extremity is stopping at the sensor on the apex. In the case of fast backward tongue motion, the snake tends to get attracted by others structures such as the floor of the mouth cavity instead of sliding along itself (see Fig. 7). The use of preprocessed images reduces the number of failures of the tracking because it emphasizes the tongue contour and suppresses the speckle patterns. On the other hand, some parts on the right side of the image (back of the tongue) are smoothed out by the preprocessing.

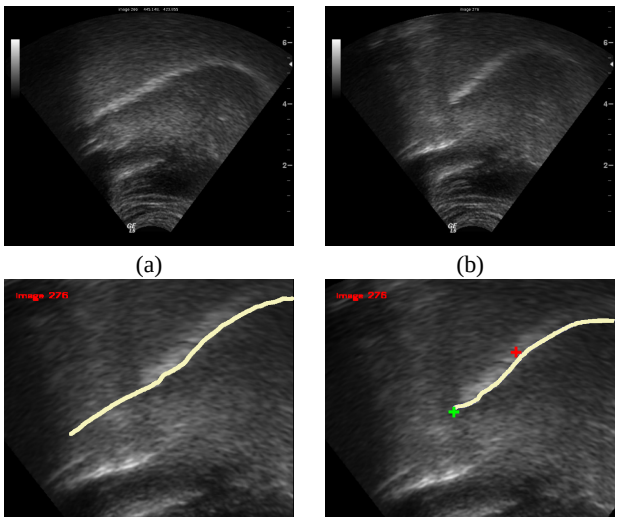


Figure 7: Effect of the sensors : (a) initialization image. (b) the tongue is moving back. (c) snake without sensors constraint (d) snake with the sensor constraint. ((c) and (d) are zooms on the apex.

5. CONCLUSION

We have presented a framework to automatically recover the tongue shape during speech processing. The first part described an acquisition system with US imaging and EM sensors, which acquires spatially calibrated and synchronized data on the tongue position during speech. The second part focused on the preprocessing of the noisy US data and presented an efficient tool to track the tongue shape by combining the estimation of the displacement, snake with constraints on the extremities, and use of EM sensors as priors to help the tracking. This method has been successfully tested on speech sequences even when fast motions occur. Residual failures were due to difficulties in imaging the back of the tongue. Future improvements will use shape priors to restrict the snake shapes to realistic ones.

REFERENCES

[1] M. Aron, N. Ferveur, E. Kerrien, M-O.Berger, and Y. Laprie. Acquisition and synchronization of multimodal articulatory data. In *Interspeech'07*, p. 1398–1401,Belgium, 2007.

[2] M. Aron, E. Kerrien, M.-O. Berger, and Y Laprie. Coupling electromagnetic sensors and ultrasound images for tongue tracking: acquisition setup and preliminary results. In *Proc. of Int. Sem. on Speech Production (ISSP'06)*, p. 435–442, 2006.

[3] M.-O. Berger, G. Winterfeldt, and J.-P. Lethor. Contour Tracking in Echocardiographic Sequences without Learning Stage: Application To the 3D Reconstruction of The Beating Left Ventricule. In *Medical Image Computing and Computer assisted Intervention, Cambridge (England)*, p. 508–515, 1999.

[4] A. Blake, R. Curwen, and A. Zisserman. A framework for spatiotemporal control in the tracking of visual contours. *Int. Journal of Comp. Vision*, 11(2):127–145, 1993.

[5] O. Engwall. Are static MRI measurements representative of dynamic speech? In *Proc. of Int. Conf. on Spoken Language Processing (ICSLP'00)*, p. 17–20, 2000.

[6] B. Horn and B. Schunck. Determining Optical Flow. *Artificial Intelligence*, 17:185–203, 1981.

[7] M. Isard and A. Blake. Contour tracking by stochastic propagation of conditional density. In *Proc. of 4th European Conference on Computer Vision, Cambridge (United Kingdom)*, volume 1064, p. 343–356, 1996.

[8] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active Contour Models. *Int. Journal of Comp. Vision*, 1:321–331, 1988.

[9] M. Li, X. Khambhamettu, and M. Stone. Automatic contour tracking in ultrasound images. *Clinical Linguistics and Phonetics*, 6(19):545–554, 2005.

[10] Min Li, Chandra Kambhamettu, and Maureen C. Stone. A level set approach for shape recovery of open contours. In *Asian Conference on Computer Vision*, p. 601–611, 2006.

[11] Hoole P. Modelling tongue configuration in german vowel production. In *Proc. 5th Int. Conf. on Spoken Language Processing (ICSLP'98)*, p. 1863–1866, 1998.

[12] M. Stone. A guide to analysing tongue motion from ultrasound images. *Clinical Linguistics and Phonetics*, 6(19):455–502, 2005.

[13] L. Vincent. Grayscale area openings and closings, their efficient implementation and applications. *1st Worksh. on Mathem. Morphology and its Applics. to Sign. Proc.*, p. 22–27, 1993.

[14] J. Weickert. Coherence-Enhancing Diffusion Filtering. *Int. Journal of Comp. Vision*, 31:111–127, 1999.

[15] D. Whalen, K. Iskarous, M. Tiede, D. Ostry, H. Lehnert-Lehouillier, E. Vatikiotis-Bateson, and D. Hailey. The haskins optically corrected ultrasound system (HOCUS). *Journal of Speech, Language, and Hearing Research*, 48(3):543–553, 2005.