Coupling electromagnetic sensors and ultrasound images for tongue tracking: acquisition set up and preliminary results*

Michael Aron[1] **, Erwan Kerrien**[1] **, Marie-Odile Berger**[1] **, Yves Laprie**[1]

[1]INRIA Lorraine - CNRS UMR7503 - Nancy-Université
615, rue du Jardin Botanique. 54602 Villers-lès-Nancy, France

{aron,kerrien,berger,laprie}@loria.fr

Abstract. *This paper describes a new method for coupling ultrasound images with tree-dimensional electromagnetic data, to recover larger parts of the tongue during speech production. The electromagnetic data is superimposed on ultrasound images after spatial and temporal calibration. Successful fusion results are presented on various speech sequences. A complete setup for evaluation of the electromagnetic system is further presented.*

1. Introduction

The development of new inversion methods and their evaluation is tightly connected to the availability of appropriate articulatory databases. Ideally, articulatory data should cover the whole vocal tract in 3D and give a time resolution sufficient for the tracking of the vocal tract kinematics (the ideal frame rate is at least 125 frames per second). In addition, these data must be available for large speech corpora and for various speakers. Therefore, the acquisition process must be fast, flexible and low cost. At present, no single imaging or sensor technique answers the above requirements alone (Engwall, 2000): Magnetic Resonance Imaging (MRI) offers a good 3D resolution of most of the vocal tract but has a very low time resolution. Ultrasound (US) provides higher temporal resolution for tongue tracking but is unable to track the apex. Electromagnetic (EM) sensors only provide sparse information on the vocal tract dynamics. As a consequence, it is necessary to register and combine several technologies to reach the objective.

Our long term objective is to provide intuitive and near-automatic tools for building a dynamic 3D model of the vocal tract from various image and sensor modalities (MRI, US, video, magnetic sensors ...). This paper focuses on the joint use of US images and EM sensors for tongue tracking. The use of EM sensors for tongue modeling has been previously advocated by many researchers, e.g. Engwall (2003). In this work, tongue modeling was achieved from MRI and the electromagnetic articulography (EMA) data were only used to correct the kinematics of the tongue. In this paper, a different and novel use of the EM data is proposed: US images are used to get the tongue contour and EM data complete the shape near the apex. Interpolation and regularization schemes can then be used to build the curve that best matches the recovered tongue contour and passes through the collected EM data. In addition, the use of a 3D EM tracker on the US transducer eliminates the need for rigidly fixing the transducer relative to the subject's head.

*Published in the 7th International Seminar on Speech Production, Ubatuba, Brazil, 12.2006

The transducer may hence be held under the chin by the subject (or an assistant), allowing for a more natural speech situation compared to e.g. the fixed Head Transducer Support System (HATS) (Stone and Davis, 1995). In order to alleviate the task of acquiring EM data, the Aurora system (Northern Digital Inc, Waterloo, Canada) is used instead of the EMA: such a system is portable, versatile (it can be used for various localization tasks and is not dedicated to speech study) and less expensive than the dedicated articulograph.

Our contributions are fourfold: (i) the Aurora technology is proved to meet the requirements of the application in terms of accuracy. (ii) The setup for coupling EM sensors and US images is presented. (iii) Spatial alignment is often achieved through manual interactive adjustments in the community and is a tedious task. Well founded and automatic methods are proposed to express EM data and US images in the same reference frame. This point is seldom addressed in the community although it is crucial to obtain a coherent dynamic model. (iv) Preliminary results are shown that exhibit fusion of tongue contours obtained from US images with EM data fixed on the tongue and especially on the apex. These results show that US and EM data are correctly coupled and can be used to recover the entire shape of the tongue above the hyoid bone.

2. Electromagnetic sensors

2.1. Materials

The tracking system under evaluation was the Aurora miniature electromagnetic system. A magnetic field generator (MFG) emits a magnetic field in a working volume (called the sensitive volume) attached to it (approx. 50 cm x 50 cm x 50 cm). Three degrees of freedom (DOF) in position and two in angulation are provided by miniature coils (0.8 mm x 8 mm). The transformation T_{em} is the translation and the rotation from the local coordinate system of the sensor to the MFG. Specifications given by the manufacturer quote a positional accuracy of 1-2 mm and an angular accuracy of 0.6 ° within the sensitive volume. A data rate of 40 Hz can be achieved with less than 6 coils plugged-in. Two 5DOF coils can be used to build a 6DOF if the two coils stay fixed relative to each other.

2.2. Accuracy and repeatability measurements

To evaluate accuracy and repeatability of the measurements, a sensor coil was fixed on a robotic arm, whose resolution is 0.013 ° for rotation and 0.48 mm for translation. We tested measurements on 3 different positions within the sensitive volume with a 5DOF sensor coil (Table 1). Position 1 was taken near the MFG (5 cm), position 2 at 30 cm and position 3 at 50 cm from the MFG. Position 2 corresponds to the approximate location of the sensors in our subsequent US/EM setup. Measurements were repeated 100 times for each position, and compared to the ground truth given by the robotic arm.

In the proposed US/EM setup, a 6DOF EM sensor coil is mounted on a US transducer (Fig.1(a)). To evaluate the influence of this device on the accuracy of the EM measurements, the same experimentation was repeated with this configuration (Table 1).

2.3. Results and discussion

Error on translation was less than 1 mm and less than 0.5 ° on rotation for positions near the MFG (1 and 2). The translation error increased significantly for position 3, with an

	5 DOF coil		5 DOF coil with an US transducer	
	Mean of translation error (in mm)	Rotation error (in degree)	Mean of translation error (in mm)	Rotation error (in degree)
Position 1	0.31	0.39	0.87	0.25
Position 2	0.53	0.50	0.76	0.20
Position 3	3.58	0.84	3.39	0.30

Table 1. accuracy of a 5DOF coil

error of 3 mm. However, accuracies correspond to the ones given by the manufacturer in Kirsch (2005) and also studied in Hummel et al. (2002). The results shows also that the accuracy decreases (loss of 0.3 mm for positions near the MFG) when the sensor is mounted on a US transducer. This loss of accuracy is due to magnetic distorsions caused by the ferromagnetic materials contained in it. Results demonstrate an adequate accuracy at position 2, which will be used in practice.

The only existing commercially available system that tracks tongue movements in three dimensions based on EM sensors is the ElectroMagnetic Articulograph (EMA – Carstens, Lenglern, Germany). The most recent model, the AG500, contains six transmitters located over a Plexiglas cube, which generate an electromagnetic field on 5DOF coils that can be glued on the tongue, the lips and the jaw. It is interesting to compare the Aurora sensors with the articulograph: Aurora coils are longer (+2 mm) but thinner than Carstens'. In terms of accuracy, the two systems seem to be equivalent. The data acquisition rate is a strong point of the AG500 systems with 200 Hz. However, the acquisition rate of the Aurora system is planned to reach 65 Hz in the near future. One important issue is the time needed to get measurements: Aurora gives them in real time, whereas the AG500 needs several hours of computation for a few minutes of speech processing. This point is important for our application, because sensors will be coupled with an other real-time modality, the ultrasound images. The flexibility, the accuracy, the time of computation, and the significantly cheaper price of the Aurora system led us to choose this material.

3. US and EM sensors for tongue tracking

3.1. Ultrasound

US is an interesting 2D image modality because it is relatively inexpensive, safe, noninvasive and can image in real-time almost any body tissue. A Logiq5 ultrasound machine (GE Healthcare, the Chalfont St. Giles, UK) was used. The transducer was a microconvex 8C, producing ultrasound signals between 5 MHz and 9 Mhz. US modality was used to get images of the tongue surface, in the mid-sagittal plane: the US transducer was fixed under the chin and held by the user. A compromise had to be found between the frequency of the US transducer, the depth of penetration, the scanning area and the image acquisition rate: the bigger the scanning area, the smaller the image acquisition rate. For the tongue surface, located between 3 cm and 7 cm from the chin during speech production, the obtained image acquisition rate was 50 Hz if the scanning area was large, and could reach more than 100 Hz for small areas (Fig. 1). With the US placed under the chin, the best US images come from sounds where the tongue surface is fairly flat and gently curved (/a/), whereas the worst images occur when the tongue has a steep slope, such as /k/, in accordance with Stone (2005).

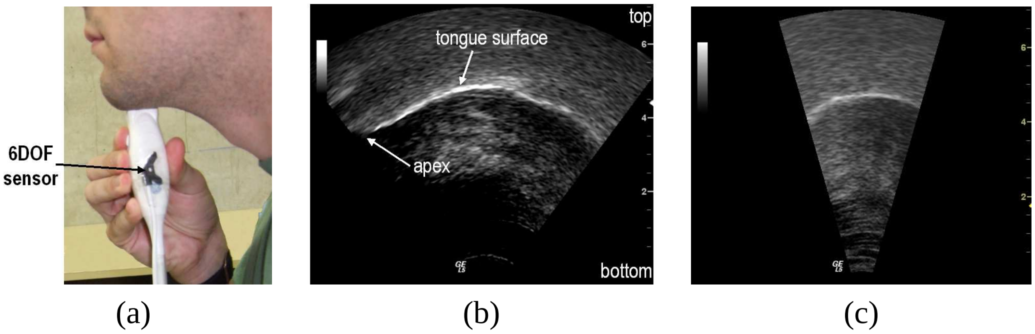


Figure 1. The tongue during /a/. (a) transducer under the chin with an EM sensor mounted on it. (b): US image at 66 Hz, approx. 6.8 cm large. (c): US image at 152 Hz, approx. 3.1 cm large.

3.2. US with EM sensor

Unfortunately, in most cases, the tip of the tongue (apex) cannot be seen on the US images, due to the air blocking the ultrasound propagation. However, the shape of the tongue can be complemented by placing one sensor on the apex. Another sensor was placed on the middle of the tongue to visually check the validity of the US/EM coupling.

To locate the EM sensors into US images, data must be expressed in the same reference frame, at any instant of time. This is made by a calibration procedure, which includes both a spatial and a temporal aspect.

4. Calibration of the US transducer and the EM system

4.1. Spatial calibration

To spatially reference EM data and US images in the same way, a 6DOF sensor was mounted on the transducer to track its motion (position and orientation) at any instant of time. These sensor data give the spatial transformation between a local reference frame linked to the 6DOF sensor and the MFG reference frame. An additional step called the spatial calibration must be added to compute the transformation T_c between the US image plane and the 6DOF sensor. Once this transformation is known, data given by the EM sensors on the tongue can be expressed in the US coordinate system.

Let $\mathcal{R}_{us}$ be a 3D coordinate system such that any pixel $p = (u, v)$ in the US image corresponds to a 3D point $P_{us} = (u, v, 0)$. Such a point is expressed in the MFG coordinate system as (see Fig. 2):

$$P_{em} = T_{em}.T_c.P_{us}$$

where T_{em} is the transformation provided by the 6DOF sensor and T_c is the searched transformation.

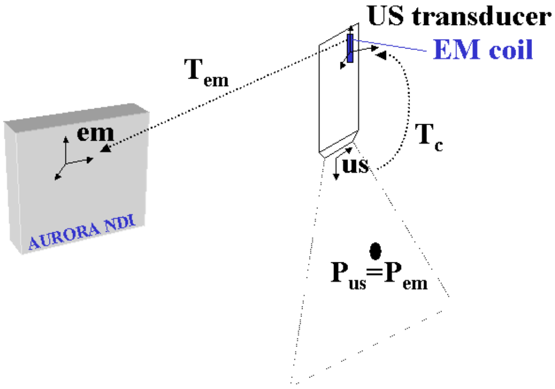


Figure 2. Coordinate system for the EM-US spatial calibration

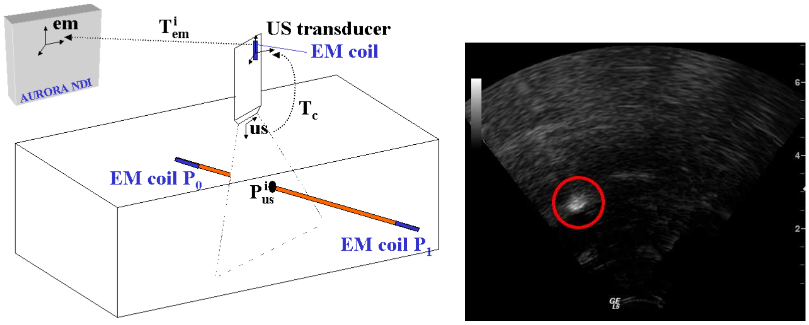
The calibration procedure consists in recovering the transformation matrix T_c (3 parameters for the translation and 3 parameters for rotation).

Note that this calibration procedure allows us to adjust for any movements of the head or the transducer (similarly to the Haskins Optically Corrected Ultrasound System HOCUS in Whalen et al., 2005), and we may hence collect the ultrasound data without a dedicated support system.

4.2. Experimental setup for the spatial calibration

A precise calibration can be obtained by scanning an object (called a phantom) with known geometric properties, and which is easily detectable in the image modality. P_{us} is detected by the user in the US image, T_{em} is given directly by the 6DOF sensor, and P_{em} is known because the object is a phantom. Therefore the transformation T_c can be computed with several P_{us} and several T_{em}. Different techniques were tested in the literature for the US/EM spatial calibration (Mercier et al., 2005), using different kind of phantoms: cross-wire with a single or multiple point target, three-wire phantoms, Z-fiducials, wall phantoms... Each design has advantages and disadvantages over the other in terms of ease of use, accuracy, and precision. There is no agreement as to which phantom design is the best.

Our phantom was inspired by Khamene and Sauer (2005): two 5DOF sensor coils were put at the extremities of a rigid wooden stick with a length of approximately 25 cm and a diameter of 3 mm, extremities forming points P_0 and P_1. This forms a line pointer, with known 3D position in the EM coordinate system and easily detectable in the US images. The pointer and the transducer were put in water at room temperature, and several images, for different positions and orientations of the transducer were acquired (30 images). In each US image, the pointer appeared as an ellipse whose center was manually selected. Such an ellipse was often larger than 10 pixels and noisy (as seen on Fig.3(b)). This issue will be worked upon in the future for improvements, but was overcame by taking a large number of calibration images. Because experimentations were made with water at ambient temperature (20 ˚ C), the speed of sound in this medium is different from the one in human tissue ($\approx 1540 m/s$). Bilaniuk and Wong (1993) showed that the speed of sound in water at 20 ˚ C is $1485 m/s$. Therefore every point in US images was corrected by shortening its depth by a ratio of $1540/1485 \approx 1.04$.



(a) (b)

Figure 3. US/EM spatial calibration (a) experimental set up for a position i of the US transducer. (b) an US image of the pointer.

The calibration problem can been seen as a minimization process, finding the 6 parameters lumped into the matrix T_c as follows (Khamene and Sauer, 2005):

$$[\widetilde{T_c}] = \underset{[T_c]}{\mathrm{argmin}} \sum_i \|(P_1 - P_0) \times (T^i_{em}.T_c.P^i_{us} - P_0)\|^2$$

where $\times$ is the cross product, and P_0 and P_1 are the extremities of the pointer expressed in the EM system. The cross product is null when the point detected in US images, once expressed in the EM system, is on the line pointer. A Powell minimization (Flannery et al., 1993) was used to minimize this criterion and find the 6 parameters.

4.3. Temporal calibration

Once EM data are spatially calibrated, data provided by the two 5DOF sensors on the tongue are expressed in the US coordinate system. The last step is to compute the delay between the starts of US and EM data acquisition for a sequence. The right delay is the one for which the second 5DOF sensor (the one not on the apex) lies on the tongue shape for the whole sequence. Practically, the distance of this sensor to the tongue was minimized as follows: the shape of the tongue was drawn on sample images in the US sequence (about 1 out of 10) by one observer. For each candidate delay, the location of the sensor in these images was determined by linear resampling. The correct delay was found to minimize the average squared distance of the sensor to the tongue.

5. Experimentation on a locutor

5.1. Protocol

The US/EM coupling system was experimented on a locutor. To isolate each sensor coil from humidity within the mouth, a plastic protection was designed and fixed by Northern Digital Inc.. The sensors were glued on a dried tongue with $Histoacryl$[1], a medical tissue adhesive used to close small wounds without suturing. They were placed on the mid-sagittal plane of the tongue, as explained in section 3.2. They remained glued on the tongue for approximately 30 minutes.

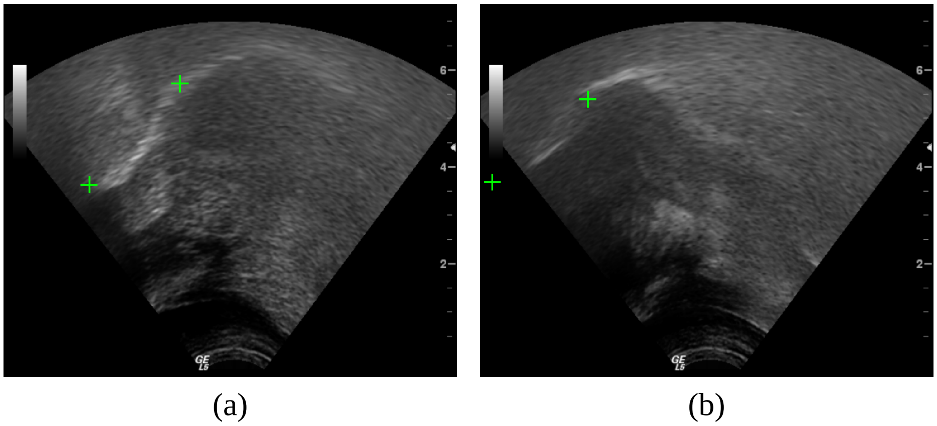
[1] http://www.bbraun.com

Four speech sequences were tested in French: ”/au/, /atu/, /aku/, /ao/, /ako/, /ae/, /ake/, /ate/” (3 times) and the sentence ”la bise et le soleil se disputaient, chacun assurant qu’il était le plus fort” (once), to see the behavior of the EM sensor during a spoken sentence at usual speed. The US frame rate acquisition was 66Hz and sensor data were linearly resampled and superimposed onto the US images.

5.2. Results

The four sequences were tested with success. First, the sensor located at the middle of the tongue appeared correctly on the surface of the tongue. This proved the spatial calibration of the section 4.1 was efficient. The sensor on the apex also moved with coherence with the tongue. Second, the temporal calibration of section 4.3 was also validated by these experimentations: the two sensors were moving according to the movements of the shape of the tongue, and the synchronization between the two modalities was satisfactory. It is interesting to note that the EM acquisition frequency (40Hz) is visually satisfactory for tongue shape imaging. This was especially visible on the sentence sequence where this movement was fast: despite the linear resampling there was no lag between US and EM data.

Images on Fig.4 show the position of the EM sensor coils (crosses) on the tongue. Fig.4(a) validates the spatial calibration of US and EM data. Fig.4(b) shows an US image where the apex is not visible and how it can be recovered from the EM data. The validation of the temporal calibration can be seen on the video sequences on our website at http://magrite.loria.fr/Confs/Issp06



(a) (b)

Figure 4. EM points (crosses) onto US tongue images: (a) /u/ from /aku/. (b) /k/ from /ake/.

6. Conclusion

This paper presents a setup for coupling EM sensors and US images in order to track and recover the complete shape of the tongue, especially the apex which is not visible on US images. It was shown that the Aurora sensors are suitable for tongue tracking with good accuracy, and the whole experimental process was explained for the US/EM coupling. This process, easily reproducible, includes a well-founded spatial and a temporal calibration which can be used for speech sequence acquisitions. It represents an alternative to

EMA, adding the shape continuity of the US modality, with only two EM sensors. For the moment, the validation of our work was only visual. Future work will focus onto validating quantitatively our results. The process of calibration will also be improved to get more accurate methods to spatially and temporally fuse data. Finally, we plan to develop our methods to track and extrapolate dynamic data of the shape of the tongue during speech, for the articulatory speech inversion processing.

7. Acknowledgment

This work is part of the ASPI project funded by the IST Programme of the Commission of the European Communities as project number IST- 2005-021324. We would like to acknowledge Mr. Sigmar Froehlich and people from Northern Digital Inc., Germany for their contribution.

References

Bilaniuk, N. and Wong, G. Speed of sound in pure water as function of temperature. *Journal of the Acoustical Society of America*, 93:1609–1612, 1993.

Engwall, O. Are static MRI measurements representative of dynamic speech? In *ICSLP-2000*, pages 17–20, 2000.

Engwall, O. Combining MRI, EMA and EPG measurements in a three-dimensional tongue model. *Speech Communication*, 41(2-3):303–329, 2003.

Flannery, B., Teukolsky, S., and Vetterling, W. *Numerical Recipes, 2nd Edition*. Cambridge University Press, 1993.

Hummel, J., Figl, M., Kollmann, C., and Bergmann, H. Evaluation of a miniature electromagnetic position tracker. *Med. Phys.*, 29(10):2205–2212, 2002.

Khamene, A. and Sauer, F. A novel phantom-less spatial and temporal ultrasound calibration method. In *MICCAI 2005*, pages 65–72, 2005.

Kirsch, S. Accuracy assessment of the electromagnetic tracking system aurora. Technical report, NDI Europe GmbH, 2005.

Mercier, L., Lango, T., Lindseth, F., and Collins, D. A review of calibration techniques for freehand 3-D ultrasound systems. *Ultrasound in Med. and Biol.*, 31(4):449–471, 2005.

Stone, M. A guide to analysing tongue motion from ultrasound images. *Clinical Linguistics and Phonetics*, 19(6-7):455–502, Sept-Nov 2005.

Stone, M. and Davis, E. A head and transducer support system for making ultrasound images of tongue/jaw movement. *Journal of the Acoustical Society of America*, 98(6): 3107–3112, 1995.

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