Multimodal acquisition of articulatory data: Geometrical and temporal registration

Michaël Aron, Marie-Odile Berger, Erwan Kerrien, Brigitte Wrobel-Dautcourt, Blaise Potard, Yves Laprie

Multimodal acquisition of articulatory data: geometrical and temporal registration

Michaël Aron
ISEN, Institut Supérieur de l'Électronique et du Numérique, Brest, France

Marie-Odile Berger, Erwan Kerrien, Brigitte Wrobel-Dautcourt, Blaise Potard, and Yves Laprie[a)]
INRIA-CNRS-Université de Lorraine, LORIA, Vandœuvre-lès-Nancy, France

(Dated: February 5, 2016)

Acquisition of dynamic articulatory data is of major importance for studying speech production. It turns out that one technique alone often is not enough to get a correct coverage of the whole vocal tract at a sufficient sampling rate. Ultrasound (US) imaging has been proposed as a good acquisition technique for the tongue surface because it offers a good temporal sampling, does not alter speech production, is cheap and widely available. However, it cannot be used alone and this paper describes a multimodal acquisition system which uses electromagnetography sensors to locate the US probe. The paper particularly focuses on the calibration of the ultrasound modality which is the key point of the system. This approach enables ultrasound data to be merged with other data. The use of the system is illustrated via an experiment consisting of measuring the minimal tongue to palate distance in order to evaluate and design Magnetic Resonance Imaging protocols well suited for the acquisition of 3D images of the vocal tract. Compared to manual registration of acquisition modalities which is often used in acquisition of articulatory data, the approach presented relies on automatic techniques well founded from geometrical and mathematical points of view.

PACS numbers: 43.70.Jt, 43.70.+i

I. INTRODUCTION

Technical advances in acquiring articulatory data have often conditioned scientific breakthroughs in speech production modeling. The work of Chiba and Kajiyama (particularly the third part entitled "The measurement of the vocal cavity and the calculation of natural frequencies") is exemplary from this point of view. It associates several modalities of imaging (X-ray photography, palatography and laryngoscopic observation of the pharynx) to determine the cross sectional area function of vowels which is then used to calculate resonance frequencies. Very substantial advances have been achieved in the acquisition of static articulatory data. By offering a millimetric accuracy, 3D MR (Magnetic Resonance) images of the vocal tract have enabled (Baer *et al.*, 1991; Story *et al.*, 1996) more accurate evaluations of vocal tract acoustic modeling. On the other hand the acquisition of dynamic geometric articulatory data still represents a challenge. Cineradiography has been abandoned in the eighties because of the health hazard due to the dose of X-ray received by subjects. Movies acquired in the past still represent a valuable source of articulatory data (Munhall *et al.*, 1995; Sock *et al.*, 2011) despite several weaknesses. By nature X-ray images are the projection of the whole head onto the image plane. The contours of the different organs are thus superimposed on the Xray image. And particularly the tongue often gives rise to several contours: that of the tongue groove in the mid-sagittal plane, and one or two others corresponding to the exterior edge approximately one centimeter left or right of the mid-sagittal plane. This explains why these contours cannot be detected easily and why contours outlined by human experts are often marred by imprecision, or even by mistakes. Additionally, the geometrical calibration is not always known precisely, and the X-ray machine used sometimes did not allow the whole vocal tract to be imaged.

These reasons explain the development of other techniques essentially based on the tracking of fleshpoints, notably X-ray microbeam and electromagnetography. X-ray microbeam (Westbury *et al.*, 1994) has been more or less abandoned partly because of the use of X-ray even if the dose is very small, as well as the cost of the machine, and above all the emergence of electromagnetography (Perkell *et al.*, 1992; Zierdt *et al.*, 1999). Beside its innocuousness, the advantage of electromagnetography is to offer a sufficiently high sampling frequency (200 Hz for the most recent machines) to analyze all speech articulatory gestures. The main weaknesses are the small number of sensors that can be tracked simultaneously, currently twelve with most recent systems but two or three have to be used to subtract head movements, and the minimal distance to respect between two sensors to avoid aberrant measures due to magnetic interferences between neighboring sensors. It is therefore possible to track three or four points on the tongue, which is enough to derive a gross approximation of the tongue shape, but not sufficiently accurate to get the precise place of articulation of many consonants. Additionally, wires connecting sensors to the articulograph change the articulation (Katz *et al.*, 2006).

The second direction of research consisted in exploiting other medical imaging techniques. Real time

[a)]Electronic address: Yves.Laprie@loria.fr

Magnetic Resonance Imaging (MRI) is probably the most interesting in the long term (Bresch *et al.*, 2008) all the more since recent technological breakthroughs (Brunner *et al.*, 2009) are likely to significantly improve the image quality in the future. But for the time being real time MRI presents strong limits for the study of speech: there are very few systems available, the spatial resolution is very poor, the supine position alters speech articulation and the machine noise induces the Lombard effect. Besides, Ultrasound (US) imaging presents some interesting advantages. It is a widely available technique, cheap, offering a good temporal sampling (between 50 and 100 Hz when imaging the vocal tract), and producing an acceptable level of acoustic noise. Stone *et al.* (1983) pioneered US imaging in the eighties and wrote a guide to analyzing tongue motion from US images (Stone, 2005). However, several technical issues have to be answered. First, the possible motion of the probe during acquisition is not taken into account. It is thus not possible to register images, and a fortiori to merge US images with images of the vocal tract acquired with another acquisition modality. Second, US images do not cover the whole vocal tract but only the mouth and the higher part of the pharynx. Third, the jaw bone and air in the possible sublingual cavity sometimes prevent visualizing the tongue.

The absence of a method allowing the probe to be localized into a head-affixed coordinate system is probably the most critical issue. Since the objective is to compensate for probe movements, one solution is to fix the probe under the chin and to prevent the subject's jaw from moving. This solution was adopted by Stone and Davis (1995) in the HATS system. It requires a strong immobilization and consequently entails a very unnatural way of articulating speech. Indeed, jaw opening is strongly reduced by the arm supporting the probe and the head is completely immobilized. Scobbie, Wrench, and van der Linden (2008) designed a helmet to hold an US probe, whose advantage is to not immobilize the head. However, arranging the helmet to the head requires tightening several screws, which probably affects the articulation of speech. Beyond the discomfort imposed to the subject, the solutions proposed above do not enable data from several recording sessions to be compared since there is no guaranty that the immobilization device is positioned exactly in the same way in all the sessions. From this point of view, Hueber *et al.* (2008) have developed an original technique to guarantee a good inter-session consistency. It consists of displaying the subject's face as it was in previous sessions and asking him to superimpose his face at best on this reference. However, this technique does not provide any metric quantification of the registration.

A less constraining way to know the US probe and the head positions is to track them along time. This solution has been chosen by Whalen *et al.* (2005) to design the HOCUS system. Infrared sensors are fixed onto the US probe, and on glasses attached with an elastic band so that they cannot move relative to the subject's head. These sensors are tracked via the Optotrack system (Northern Digital inc.). Actually, fleshpoints behind the ears are probably less mobile but the nature of the Optotrack system which utilizes infrared emitting diodes (IREDs), requires the sensors be visible from the cameras. The designers of the HOCUS system preferred not to add probe immobilization device so as to avoid spurious articulatory compensation gestures. Therefore, there is no guarantee that the probe remains in the mediosagittal plane. This setup gives the position and orientation of the probe and head in the optical coordinate system. However, the plane of the US image cannot be known. Designers thus added three sensors onto the US probe so that the plane defined by these points approximately coincides with that of the US image. The geometrical transformation between sensors glued on the probe and the US image is not calculated but only estimated by hand. This lack of geometric calibration results in an uncontrolled inaccuracy.

Beyond this inaccuracy, this system does not allow for a point detected in the US image to be located in 3D. Indeed, the Optotrack system is only aware of the position of the IREDs, whereas the point is detected with respect to the origin pixel of the image. The resolution of the US machine, as provided by the manufacturer, helps to translate this location in millimeters but the distance of at least one IRED to the origin pixel of the image is further required to locate the point in 3D.

Moreover, the tongue is not completely imaged by the US transducer. The hyoid bone hides the tongue root in the lower pharynx and the US beam is generally not sufficiently wide to cover the whole tongue. In the front part of the mouth, the tongue tip is often not visible for two reasons. First, the US beam is reflected by the jaw bone and thus does not reach the tongue tip. Second, the mouth floor of the sublingual cavity, when it exists (for certain articulations, like /ʃu/ for instance), is the first interface with air reached by the US beam, and prevents the tongue tip from being imaged.

The present paper describes the ARTIS system which incorporates an US machine to capture the tongue contour, an electromagnetic (EM) localization device to locate the US probe and the tongue tip via EM sensors, and finally stereo vision cameras to track markers painted on the speaker's face. All these acquired data are synchronized with the recorded audio track.

First, we describe the complete acquisition setup and present experiments intended to evaluate the geometric accuracy of the system. Then, we present the registration of US data with other imaging modalities, which resorts to the calibration of the US machine with respect to the EM system. Thereby, the geometrical transformation between the coordinate systems of these two modalities is determined. This calibration enables the location of any 2D point of the US image to be known in the EM sensors 3D coordinate system. It avoids resorting to manual registration which depends on the human experimenter's experience and whose validity and accuracy cannot be assessed easily. Practically, the real time registration offers more flexibility to experimenters and subjects since no immobilization is required to keep the head perfectly still.

Finally, we show how this system can be used and how other data can be merged.

II. DESCRIPTION OF THE SYSTEM AND OVERALL PRINCIPLE

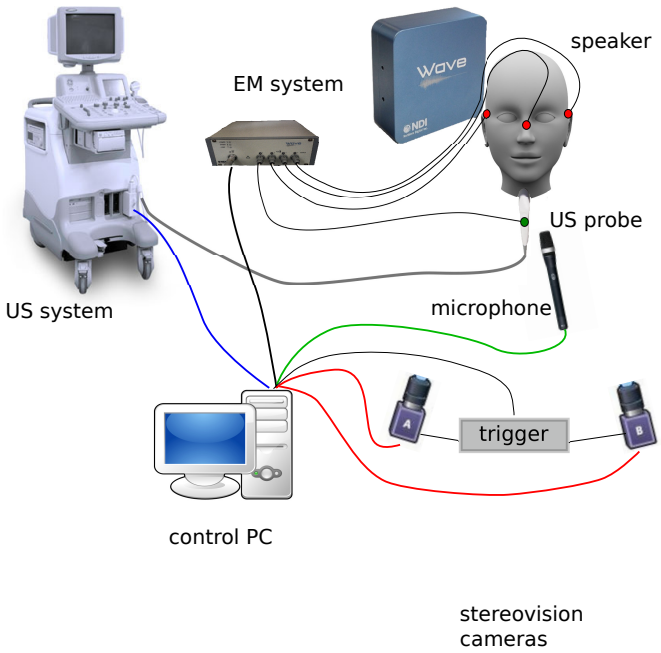


FIG. 1: Architecture of the acquisition system.

Fig. 1 and 3 present the overall system and setup. Each acquisition has its own coordinate system and its own temporal scale. The objective is to express all the data acquired within a unique coordinate system and temporal scale. The reference coordinate system is that of the EM acquisition device because it enables the US probe to be tracked and the head movements to be compensated for, but eventually the objective is to express all data in the coordinate system of the speaker's head. The reference temporal scale is that of the internal clock of the PC which is a specialized chip with a precision of the order of magnitude of one nanosecond. Each captured event is time stamped.

Expressing all the geometric data in one reference coordinate system covers three issues:

1. determining the three dimensional position of one pixel for each image modality within the coordinate system attached to this modality. This concerns US imaging and optical cameras. It consists of calibrating these two modalities. This refers to the determination of their intrinsic geometric parameters so as to compute the three dimensional position of points imaged or tracked in the coordinate system. In practice this requires a ground truth to be provided by an object whose geometric properties are known and which can be imaged by the modality to be calibrated.

2. determining the three dimensional position of one pixel of the US modality within the reference coordinate system, i.e. that of the EM localization device. This is achieved by fixing sensors on the US probe to track it. Additionally, gluing sensors behind the subject's ears and on nose bridge enables the position of the head to be tracked. It is thus possible to determine the position of US pixels within a coordinate system attached to the speaker's head.

3. merging other modalities, which can be recorded at the same time, such as images of the speaker's face, or prior MRI images of the speaker's vocal tract to get the palate shape for instance. In this case the solution consists of computing the Euclidian transformation between two clouds of points of the upper part of the speaker's face, each cloud being provided by a modality.

The first component of the ARTIS system (see Fig. 1) is a Logiq5 US system (GE Healthcare, the Chalfont St. Giles, UK). Before designing the system, we assessed several portable or fixed US machines and retained the Logiq5 because it yields images of a better quality than those acquired with a portable system, while being at a reasonable cost. This choice made in 2009 could probably be questioned today with the emergence of more effective portable systems. US movies are directly acquired into the memory of the Logiq5 in the form of DICOM (Digital Imaging and COmmunication in Medicine is the standard file format in the domain of medical imaging) files. Files are then transferred and decoded to get the images. The utilization of DICOM files instead of the US video output presents the advantage of keeping the source sampling frequency and the image quality of the US machine, instead of that imposed by PAL video encoding system which lowers the sampling frequency. We chose a microconvex 8C transducer, producing US signals between 5 MHz and 9 Mhz. It offers a good coverage of the tongue and its small size and curved shape make it comfortable for the subject. The sampling frequency is set to 66Hz.

The second component is the Wave miniature EM system (Northern Digital inc., see Fig.2) including a magnetic field generator (MFG), a control unit and up to twelve miniature coils (2 mm x 3 mm) tracked at a sampling frequency of 100 Hz and providing 5 degrees of freedom (DOF) data: the position (3 DOF) and orientation (2 DOF) of a coil are expressed in the coordinate system of the magnetic field generator.

One 6 DOF EM sensor can be built using two 5 DOF EM sensors, rigidly fixed relative to each other. We used such a 6 DOF sensor, called MagTrax sensor, glued onto the US probe to track it (see Fig. 4.b). The 6 degrees of freedom define a full rigid body transformation (the 3 angles of the rotation and the 3 coordinates of the translation), called T_{em} which can be represented in the form of a 3x4 matrix (the left 3x3 submatrix is the rotation and the right column vector is the translation). It represents the transformation between the local coordinate system rigidly linked to the MagTrax sensor, i.e. the US probe, and the coordinate system of the MFG, which is fixed in the room. For reasons of calibration and accuracy, the manufacturer limits the working volume (called

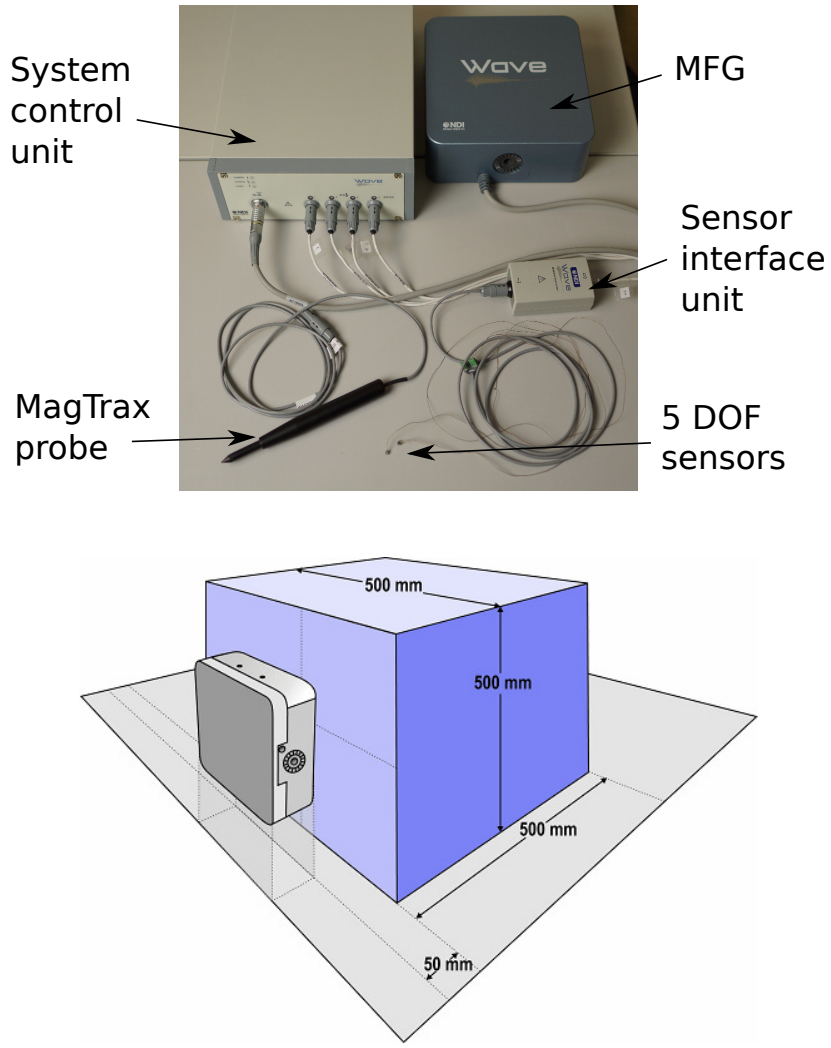


FIG. 2: Wave system and measurement volume with the magnetic field generator.

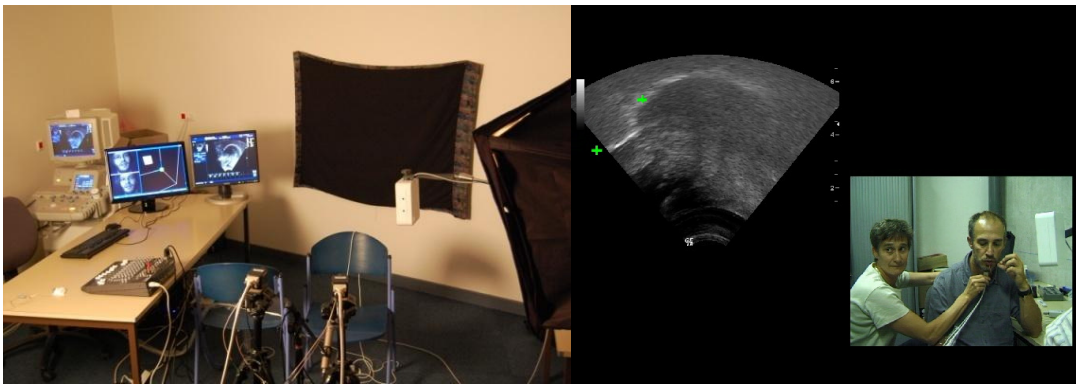


FIG. 3: Complete acquisition system (left) and example of acquisition (right).

the sensitive volume, see Fig. 2 top) to a cube of edge length 50 cm.

The third component is a pair of two synchronized JAI-Pulnix TM-6740CL cameras. These cameras are configured to acquire 640x480 grayscale images at 198 frames per second. They are synchronized via an external trigger (CC320 Machine Vision Trigger Timing Controller, Gardasoft). Two Super Cool-Lite 9 (interFit Lighting) projectors were used to reduce the flickering observed due to neon lights.

Markers painted onto the speaker's face are used either for studying speech production (those painted onto

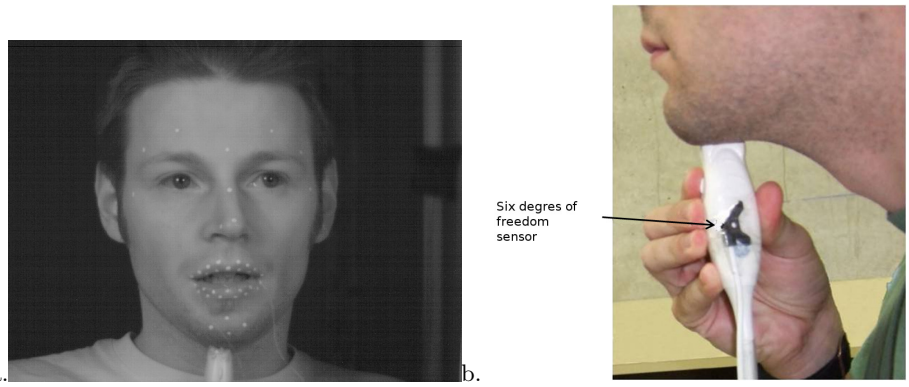


FIG. 4: (a) Subject and sensors painted onto his face. (b) Subject and the US probe with the six DOF sensor.

lips and chin), or for compensating for head movements (those painted onto the forehead and the nose edge).

The different modalities, including the audio recording system, are supervised by a control PC which uses four RAID0 disks to achieve a transfer speed around 360 Mo/s. The acquisition software was developed with Microsoft Visual Studio 2008 in C++. In addition to offering a simple access to the acquisition libraries (the acquisition and control libraries Sapera LT SDK to capture images with the two cameras, and the audio component of the DirectSound of the DirectX API to record the speech signal), it also enables a direct access to low level Windows primitives (CreateFile/WriteFile API) which turned out to be necessary to benefit from the full potential I/O transfer rates of the RAID disks. A complementary module used for displaying the orientation of the US probe is developed in Python (see section IV.A.3).

Fig. 3 presents the room prepared for acquisition and an acquisition. Fig. 4 presents the subject ready for an acquisition (with markers painted onto the face, EM sensors, and the US probe).

III. GEOMETRICAL PRECISION OF THE WAVE SYSTEM

Specifications given by the manufacturer quote a positional accuracy estimate of 1 to 2 mm and an angular accuracy of 0.6° within the sensitive volume (see Fig. 2) for data acquired at 100 Hz simultaneously with up to 6 coils.

We first checked the capacity of the Wave system to provide accurate data, in the conditions met in our setup. To evaluate accuracy and repeatability of the measurements, a sensor coil was fixed on a robotic arm, whose resolution is 0.013° on rotation and 0.48 mm on translation. We tested 3 different positions within the sensitive volume with a 5DOF sensor coil (Table I). The first position was taken near the MFG (5 cm), the second at 30 cm, and the third position at 50 cm from the MFG. The second position corresponds to the approximate location of the sensors in our subsequent acquisition setup. Measurements were repeated 100 times for each position, and compared to the ground truth given by the robotic arm. In the setup of our acquisition system, the US transducer is tracked by using a 6 DOF sensor (the MagTrax sensor, see Fig. 5.a). Its tracking may be affected by EM disturbances that the US transducer may cause. We thus repeated the same experiment with a 5 DOF sensor fixed onto the transducer (Table I).

Table I presents the results obtained. Errors on translation were less than 1 mm and less than 0.5° on rotation for positions near the MFG, i.e. the first two positions (5 cm and 30 cm). The translation error significantly increases for the third position (50 cm), with an error of 3 mm. These accuracies correspond to the ones given by the manufacturer in (Kirsch, 2005) and also studied in (Hummel *et al.*, 2002). The results also showed that the accuracy decreased of about 0.3 mm for positions near the MFG when the sensor was mounted on the US transducer. This loss of accuracy is due to magnetic distortions caused by the ferromagnetic metals contained in it. However, results demonstrated an adequate accuracy at the second position, i.e. where the speaker's head is located when investigating speech production. Sensors can thus be used either as pointers or reference points to achieve the geometrical registration between the different modalities.

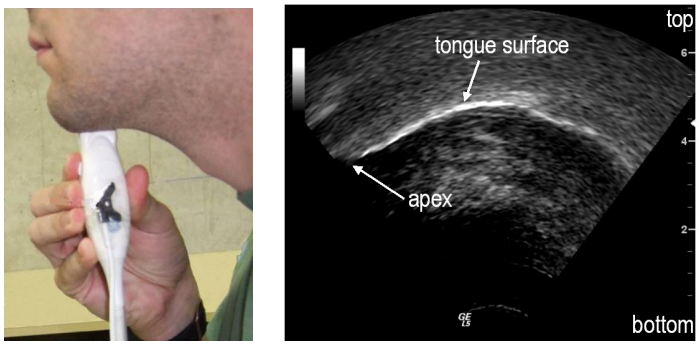
IV. COUPLING US AND EM DATA

The smaller area scanned by the US signal, the faster the image acquisition rate. It is thus necessary to find a compromise between the frequency of the US signal, the image depth, the scanning area and the image acquisition rate. Since the tongue is located between 3 and 7 cm from the transducer during speech production, the image acquisition rate is between 50 Hz for a wide coverage and 100 Hz to observe narrow regions (Fig. 5).

We chose an acquisition frequency of 66 Hz, an image size of 532x434 pixels, a resolution 0.17 mm per pixel, and a depth of 8 cm. These settings seem to us the best compromise between the possibilities of the US machine and the imaging of tongue contours (see Fig. 5.b).

	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 I: Accuracy of a 5 DOF coil. Distances to the MFG are: 5 cm (Position 1), 30 cm (Position 2) and 50 cm (Position 3).



(a) (b)

FIG. 5: (a) position of the transducer with an EM sensor fixed on it. (b): US image acquired at 66 Hz, approx. 6.8 cm large. (c): US image acquired at 152 Hz, approx. 3.1 cm large.

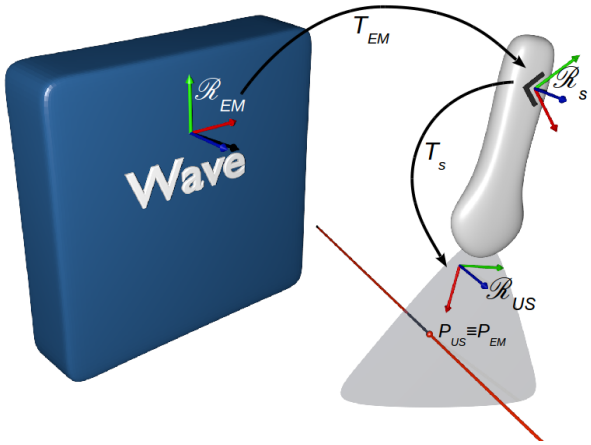


FIG. 6: Coordinate systems for the calibration of US imaging with respect to the EM system and transformations used.

A. Calibration of the US transducer and the EM system

All data must be expressed in the same coordinate system before they can be coupled. Since all EM data, and in particular those tracking the head, are expressed in a coordinate system fixed with respect to the MFG, we used this frame as reference. In order to track the motion of the US probe, a MagTrax 6DOF sensor was mounted on the probe. Fig. 6 depicts the geometry of the system and the different coordinate frames involved: $\mathcal{R}_{EM}$ is the frame attached to the MFG (reference frame); $\mathcal{R}_s$ is the frame attached to the 6DOF sensor; and $\mathcal{R}_{US}$ is the frame attached to the US image. A pixel site with coordinates (c, r) (column index, row index) corresponds to the 3D point $P_{US} = (c, r, 0)$ in $\mathcal{R}_{US}$. This 3D point is expressed as P_{EM}, in $\mathcal{R}_{EM}$, such that:

$$P_{US} = S_{US}^{-1}.T_s.T_{EM}^{-1}.P_{EM} \quad (1)$$

where . means the application of a transform: T_{EM} is the rigid transform between $\mathcal{R}_s$ and $\mathcal{R}_{EM}$, it is provided by the sensor as explained in section II and varies when the sensor moves; T_s is the rigid transform between $\mathcal{R}_s$ and $\mathcal{R}_{US}$: it is unknown but fixed; $S_{US} = \text{diag}(r_h, r_v, 1)$ is the scaling matrix whose diagonal elements are the horizontal (r_h) and vertical (r_v) pixel dimensions.

The calibration procedure consists in recovering the transformation matrix T_s (3 parameters for the translation and 3 parameters for rotation) as well as both pixel dimensions (r_h, r_v).

1. *Experimental setup for calibration*

The calibration of an imaging modality with respect to another is generally obtained by considering features, e.g. points, visible in both modalities. In our case, the direct calibration of both modalities is impossible because EM sensors are invisible in US images. An object, called phantom, must thus be designed with known geometrical properties and features easily detectable in both modalities. Different techniques were tested in the literature for the US/EM spatial calibration (Mercier *et al.*, 2005), using different kinds of phantoms: cross-wire with a single or multiple point targets, three-wire phantoms, Z-fiducials, wall phantoms... Each design has advantages and disadvantages in terms of ease of use, accuracy, and precision. There is no agreement about the best phantom design. We chose line features as a good compromise to tackle two difficulties: imaging the features with the US probe, and automatically detecting the features in the images. We designed our own phantom, based on the US phantom proposed by CIRS Inc for 2D evaluation (Model 555, CIRS Inc., Norfolf, VA). The target lines

are wires immerged in a background medium calibrated to mimic the US characteristics of human tissues (in particular, speed of sound of 1540m/s). However, we had to complete this basic design in three ways.

First, the model by CIRS Inc. presents one beam of parallel wires. This enables the whole transform parameter set to be retrieved, but for a translation along the common wires direction. Therefore, a second beam of parallel wires was added, orthogonal to the first one.

Second, each line equation is known, by design, in a coordinate frame attached to the phantom. Since the wires had to be localized in $\mathcal{R}_{EM}$ (Khamene and Sauer, 2005), each wire extremity was marked onto the phantom casing. A MagTrax probe (see Fig. 2) was used to localize these points.

Third, each beam was designed to present a different and asymmetric pattern in the images to enable automatic labelling of the lines.

Fig. 7 gives all the characteristics of our phantom. It was custom built by CIRS Inc., Norfolk, VA.

2. Workflow for the calibration

Equation 1 relates frames $\mathcal{R}_{EM}$ and $\mathcal{R}_{US}$. However, the phantom wire locations are specified within a phantom-attached coordinate frame $\mathcal{R}_p$. Wires could easily be localized in $\mathcal{R}_{EM}$ by pointing the MagTrax probe at each wire extremity, visible on the phantom casing. But the EM measurement noise, combined with the manual pointing inaccuracy hamper the use of those locations. Therefore, we rather first determined the rigid transform T_p that related $\mathcal{R}_p$ to $\mathcal{R}_{EM}$. And in a second step, the transform $T_c = S_{US}.T_s$ was calibrated.

a. Calibrating T_p: Once the phantom had been placed in the measurement volume of the MFG, each wire extremity was localized in $\mathcal{R}_{EM}$ with the MagTrax probe: the positions acquired during a 2 seconds acquisition each were averaged to reduce the influence of noise. Then the rigid transform T_p was determined by minimizing the following criterion:

$$\mathcal{C}_p = \frac{1}{N_w} \sum_i d(\Delta_i, T_p.P_{i,0})^2 + d(\Delta_i, T_p.P_{i,1})^2$$

where N_w is the number of wires, $d(.)$ is the point to line Euclidean distance, $P_{i,0}$ and $P_{i,1}$ are both extremities of line Δ_i, in $\mathcal{R}_{EM}$, and Δ_i are the equations of the phantom lines known by design in $\mathcal{R}_p$.

A Powell minimization (Flannery *et al.*, 1992, Chapter 10) was used to minimize this criterion and find the 6 parameters of T_p. We found a residue $\sqrt{\mathcal{C}_p} = 0.65$ mm which is compatible with the precision of the EM sensors. As a result, each target line could be expressed in $\mathcal{R}_{EM}$ by applying T_p.

b. Calibrating $T_c = S_{US}.T_s$: The transform T_c was determined in two steps:

- **Initialization:** An US image was taken showing only the beam along the Z direction. Since those wires are parallel, there exists an affine transform between the pattern as seen in the US image, and the pattern shown on Fig. 7 (left). Four wires were manually identified in the image (see Fig. 8 (left)) to estimate this transform. This transform provides estimates for the pixel dimensions r_h^0 and r_v^0, as well as the in plane rotation and translation. Assuming the US image plane is orthogonal to the Z direction in $\mathcal{R}_p$, i.e. there is no out-of-plane rotation, and leaving out the unknown out-of-plane translation, provides a complete rigid transform T that relates $\mathcal{R}_p$ to $\mathcal{R}_{US}$. T_c is initialized as:

$$T_c^0 = S_{US}^0.T.T_p^{-1}.T_{EM}^{-1} \quad (2)$$

  where T_{EM} is given by the probe-attached MagTrax sensor.

- **Optimization:** Various US images are taken with the tracked probe, in diverse orientations, including images showing both orthogonal wire beams. A strong bilateral noise reduction filter (window size=10 pixels, spatial standard deviation=10 pixels, followed by a threshold are applied on each image. The wire markers are extracted as the center of gravity of the resulting spots (connected components) and automatically labelled with wire indices.

  Transform T_c is determined as minimizing the following criterion:

$$\mathcal{C}_c = \frac{1}{N_m} \sum_j \sum_{i \in \mathcal{L}_j^*} d(Q_{i,j} - \Pi.T_s.T_{EM,j}.T_p.\Delta_i)^2$$

  where $Q_{i,j}$ is the detected marker corresponding to ith wire in the jth image, $\mathcal{L}_j^*$ is the index set of lines that intersect the image plane in the jth image and matches a detected marker in the US image, N_m is the total number of matches in all images, Π is an operator that computes the intersection of a line with the plane $z = 0$, and $d(.)$ is the 2D Euclidean distance. Note that T_{EM}, given by the probe-attached sensor, depends on the image and is therefore indexed by j. Since our initialization procedure proved to be good enough, a Powell minimization was used to determine the transform T_c (8 parameters). Fig. 8 (right) provides a sample calibration result. We found a residue $\sqrt{\mathcal{C}_c} = 8.861$ pixels. The US image resolution was estimated in the process and found to be 0.174 mm in the horizontal direction and 0.184 mm in the vertical direction. The significant difference with the figures provided by the US system highlights the need for calibrating the image resolution. Moreover, the anisotropy of the pixel size cannot be neglected and has a significant impact on the registration accuracy.

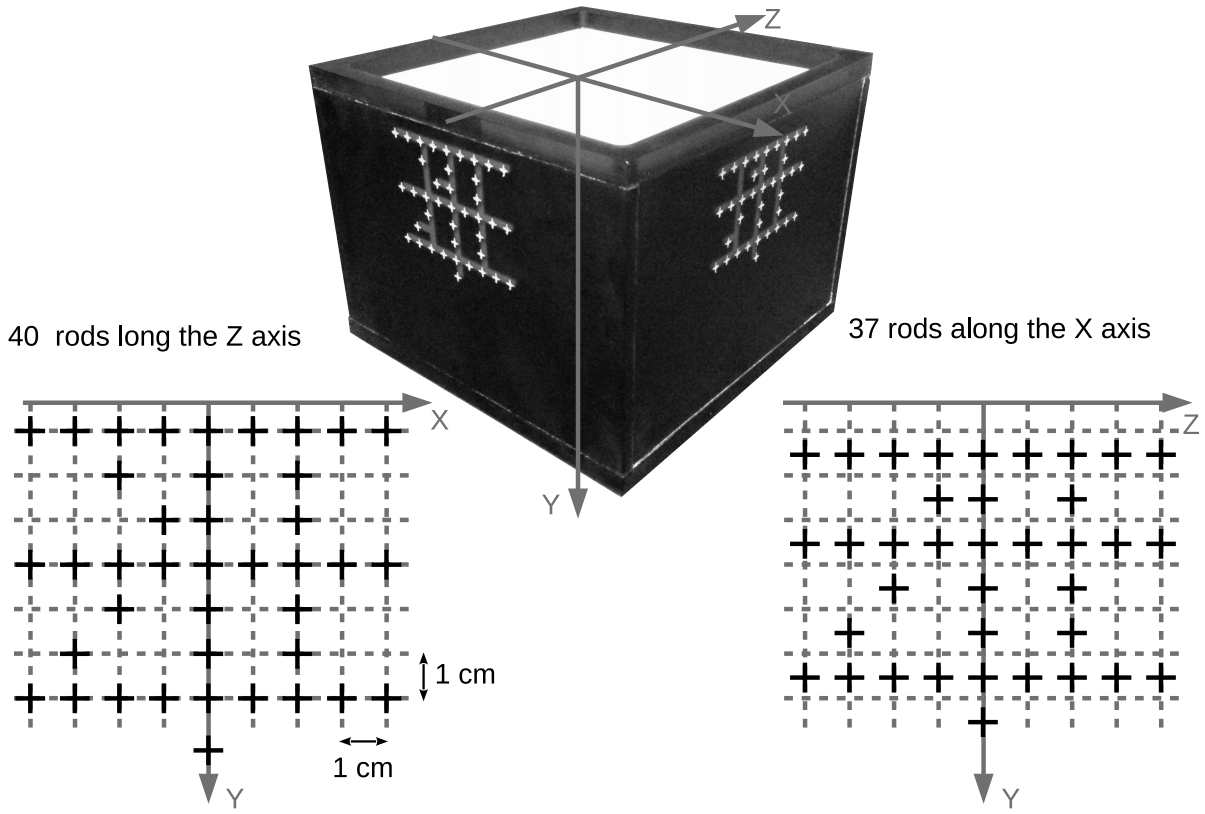


FIG. 7: Specifications of our phantom designed for US/EM spatial calibration. The black crosses on the left give the extremities for the set of wires parallel to the Z direction. The basic spacing is 1cm. The black crosses on the right give the extremities of the second set of wires parallel to the X direction, they are intertwined with the first set with a shift of 0.5 cm along the Y direction. The photo in the middle displays the coordinate axes and shows the wire extremities of both sets marked in white on all four faces of the phantom casing.

3. Direction of the US probe

The geometrical calibration enables the positions of the US data to be known in the EM coordinate system. However, the US probe has to be oriented by the experimenter in the subject's mid-sagittal plane to capture relevant articulatory data. During preliminary acquisitions, we noticed that some US sequences were not correct because the plane of the transducer did not contain the sensors glued on the tongue and assumed to lie in the mid-sagittal plane. To cope with this problem, we developed a graphical interface displaying the respective positions of all the sensors as well as the US plane. When the sensors on the tongue are too far from the US plane, a red warning is displayed to ask the experimenter to correct the probe position.

V. SYNCHRONIZATION

All the modalities of our system are linked to a control PC which controls the recording process of each modality. The setup between all the different devices used for the speech acquisition is summarized on Fig. 1. The main recording characteristics of each modality are summarized in Table II.

When dynamic data are recorded with several modalities, a synchronization process is required to align them on the same temporal scale. This implies that the acquisition time of data acquired with each modality is expressed into a common reference time frame. This process must be carefully achieved because a temporal misalignment error, i.e. a spurious delay between data may be disastrous. For example, an error of 2% on the acquisition frequency of audio data (43.2 kHz instead of the theoretical value of 44.1 kHz) generates a shift of 300 milliseconds on audio data after 15 seconds of acquisition.

Despite its importance, the problem of synchronization in multimodal recording systems is generally underestimated. In (Whalen *et al.*, 2005) and in (Stone, 2005), a video camera recorder was used to synchronize sound with US images (by downsampling US images at 30 Hz, i.e. the frequency of the NTSC format). Stone (2005) observed a variable delay of up to several seconds between sound and US images on the output sequences. In the articulograph system, an external trigger is used but Qin and Carreira-Perpiñán (2007) have noticed a delay of 15 milliseconds between audio and EM data. Because the source of the synchronization error is hardly identifiable (this could be a delay between the acquisi-

	EM	US	stereo vision	audio
Frequency	100 Hz	66 Hz	200 Hz	44100 Hz
Recording time	unlimited	15 seconds	unlimited	unlimited
Data format	TXT files	DICOM	PGM images	WAV files
Recording process	Real time	Record button	Real time	Real time

TABLE II: Main characteristics of the modalities.

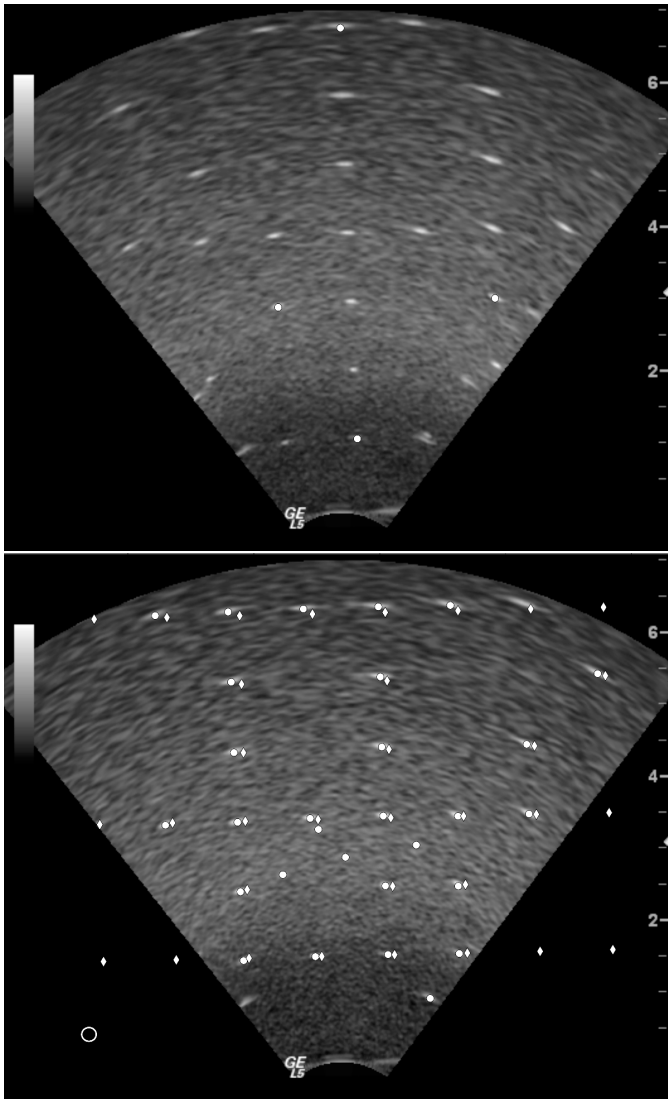


FIG. 8: EM/US calibration: (left) manual identification of 4 points to initialize T_c (disks); (right) a sample result after iterative optimization: automatically extracted wire markers (disks), the position of the wires after optimization (diamonds), the circle in the bottom left of the image has a radius of 1 mm.

tion and the recording of the data, a bad acquisition frequency given by the manufacturer, the time required by the synchronization material to process the data...), the synchronization of the multimodal data is often not well addressed in the existing acquisition systems.

A. Experimental synchronization process

All modalities are recorded by the control PC and each of the recorded data is time stamped according to the time of the control PC considered as the reference time frame. The time stamp is redundant and contains the time elapsed since the acquisition started given by the internal PC clock which has a very high precision, and the number of audio samples recorded. This double stamping allows the inaccuracy of the audio sampling frequency to be compensated for. Recording is thus completely independent from the sampling frequencies of all the acquisition devices used by the system.

The delays between the acoustic signal acquisition and each other acquisition modality are then measured experimentally. The synchronization thus amounts to subtract each of these delays to the corresponding measure.

We now describe the experimental process for computing the delays between all the modalities and the reference time frame. An event, easily identifiable in both the acoustic and the target modalities, is generated. In the reference time frame, timestamps of this event are identified for each modality. Their difference estimates the delay between both modalities. This procedure was repeated several times to obtain an evaluation of the standard deviation of the delay so as to check its stability across several acquisitions. Since all data of the modalities could be time stamped by the control PC, the estimation of the acquisition frequency was not required.

1. Audio-Electromagnetic synchronization

Fig. 9 shows the experimental setup for the audio/EM synchronization. A user hit the microphone with the MagTrax probe. This events was visible on both the EM signal (the movement of the EM pointer stops) and on the audio signal (peak).

We found a mean delay of 67.8 milliseconds with a standard deviation of 8.9 milliseconds for 20 experiments. Since the standard deviation was less than half the acquisition period of the EM system, the variability of the delay was considered as negligible. But the mean value of the delay shows that there was a difference of 3 EM samples between an audio sample and its corresponding EM sample.

FIG. 9: Experimental protocol for audio/EM synchronization: an EM pointer hits the surface of a microphone.

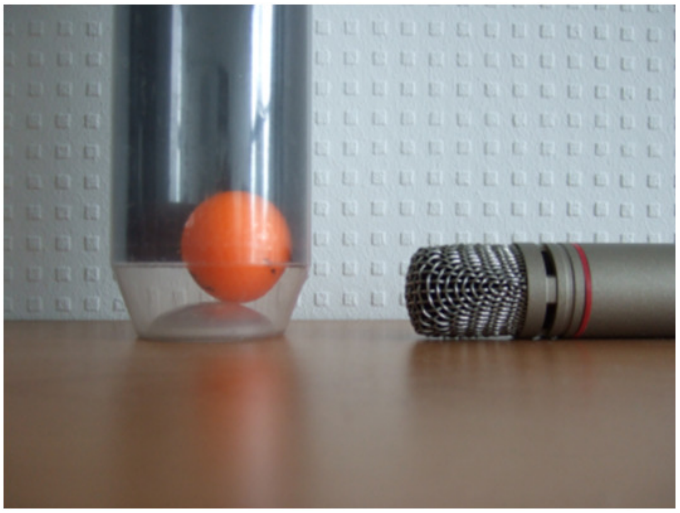


FIG. 10: Experimental protocol for audio/video synchronization: a ball hits the bottom of a plastic bottle. The scene is recorded with a stereo vision system.

2. Audio-Stereo vision synchronization

Fig. 10 shows the experimental setup for the audio/video synchronization. The event was a ball hitting the bottom of a plastic bottle. This event was both visible on the video images and on the audio signal.

The mean delay was 9.78 milliseconds with a standard deviation of 0.23 milliseconds for 20 experiments. Since the standard deviation of the delay was less than half the acquisition period of the video system, its variability was considered as negligible. The mean value of the delay showed that there was a difference of 2 video samples between an audio sample and its corresponding video sample.

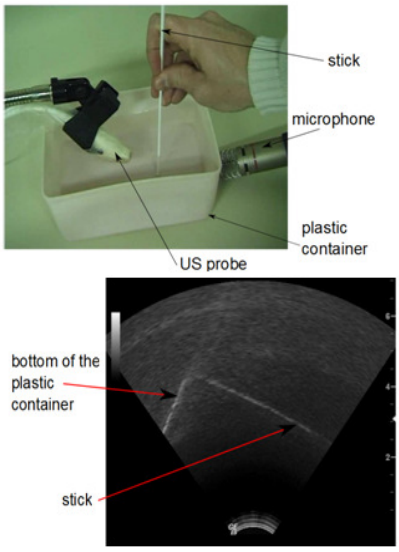


FIG. 11: Experimental protocol for audio/US synchronization: a stick hits the bottom of a plastic container (Fig. 5.a). This event recorded by the microphone is also visible in the US image (Fig. 5.b).

3. Audio-ultrasound synchronization

Our US machine, as many commercial ones, has two particularities: the duration of the recording is limited (15 seconds whatever the acquisition frequency) and the recording process is launched by pressing a recording button which saves the last 15 seconds on the US hard disk. In our system, a pulse is sent from the control PC to the US system, simulating a push on the recording button of the US system at the end of the acquisition. Technically, we built a cable plugged into the serial port of the control PC and linked to the recording button of the US system. For this reason, each US image cannot be time stamped by the control PC and only the pulse triggering the recording is visible in the reference time frame of the control PC. Therefore, in addition to the estimation of the delay between sending the pulse and recording the last US image, an estimation of the acquisition frequency of the US system must be calculated. This enables each US image received by the control PC to be time stamped. Fig. 11 shows the experimental setup for the audio/US synchronization. The event was the impact of a stick with the bottom of a plastic container full of water. This event was visible both on the US images and on the audio signal. This event was repeated several times during a same US sequence of 15 seconds.

The mean delay was 14.9 milliseconds with a standard deviation of 8.2 milliseconds for 20 experiments. Since the standard deviation is more than half of the acquisition period of the US system (15.1 ms), the delay variability corresponds to +- 1 US image. These experiments also enabled the sampling frequency to be evaluated at 65.92 Hz with a standard deviation of 0.02 Hz. Even if the standard deviation is very low, meaning that the acquisition frequency is stable, a frequency slightly

lower than that provided by the manufacturer (66 Hz) has been found. This difference is not negligible after several seconds of US acquisition (1 US sample after 15 seconds), which proves that the determination of the frequency is necessary.

VI. MERGING US DATA WITH STEREO VISION AND MRI DATA

In many cases, for instance the creation of a full talking head model, it is necessary to use other modalities apart from US images, as they only allow for the tongue to be visualized. The speaker's face and the fixed vocal tract wall (mostly composed of the hard palate) were acquired using other modalities, and all these need to be merged together. Geometrical face measures are acquired by means of the two synchronized cameras forming the stereo vision system (Wrobel-Dautcourt *et al.*, 2005), which provides the 3D positions of markers painted on the speaker's face. Similarly to the US probe, stereo vision has been calibrated via classical calibration methods used in computer vision. For practical reasons (essentially the ferromagnetic nature of the calibration target) it was not possible to compute the geometrical transformation between stereo vision and the EM system, and thus to register these two modalities directly.

Of course, it is unimaginable to use EM sensors inside the MRI machine used to determine the hard palate geometry and it is necessary to call on another technique. The solution consists of registering the cloud of face data measured with any two systems (stereo vision, electromagnetography, magnetic resonance imaging) by computing the rigid transformation between these two point clouds via the iterative closest point algorithm (Besl and McKay, 1992), which is intensively exploited in computer vision.

The clouds of points correspond to markers painted on the speaker's face for stereo vision, to points measured by scanning the speaker's face with a MagTrax probe, and to the face surface extracted from a 3D MRI acquisition of the speaker's head (see Fig. 12).

From a practical point of view this algorithm works best when one of the two clouds of points corresponds to a fairly fine mesh. We thus used a high resolution scan of the face of the speaker (acquired with a 3D InSpeck system) as a common reference surface. The ICP algorithm is thus applied between one cloud of points (stereo vision, electromagnetography, MRI) and the digitized face surface. The four clouds of points or surface correspond to the face at a rest position.

This registration strategy enables US images to be merged with stereo vision or MRI data via the EM system. We exploited the latter possibility in the experiments presented in the next section.

VII. VALIDATING EXPERIMENT AND EXAMPLE OF MEASUREMENT WITH THE SYSTEM

A. Impact of tracking the US probe via the EM system

The first experiment was intended to illustrate the positive impact of tracking the US probe along the acquisition versus the strategy of keeping the probe position measured in the first image of the sequence. The speaker's face and the US image had been registered together as explained in the previous section.

Fig. 13 illustrates the tracking vs non-tracking impact on four images extracted from a sentence. The left images were obtained by taking into account the variability of the probe position during the acquisition given by the sensor glued on it whereas the right images did not take into account the probe position. In addition to the sensors used to track the speaker's head and the US probe, two sensors were glued on the tongue (one on the tongue blade and a second close to the tongue tip). It turned out that both sensors remained in contact with the tongue as expected in the tracking strategy (left images) whereas they substantially deviated from the tongue in the second case. This clearly shows the importance of tracking the US probe during the acquisition.

B. Measuring the tongue to palate distance by merging ultrasound and MRI data

Evaluating the positions of the speech articulators is crucial for a better understanding of articulatory gestures and speech production. Of particular interest is the tongue to palate distance, and the place where it is minimal. This measure is important since it can be related to the acoustic properties of the speech sounds. However, it cannot be measured directly in normal phonatory conditions. Indeed, US only provides information on the tongue surface alone, and in the other hand MRI provides insufficient sampling frequency, triggers Lombard effect due to the strong environment noise, and could also alter speech articulation due to the supine position.

In the second experiment reported here, we used this measure to investigate and design an MRI protocol intended to acquire one hundred high definition static 3D MRI images of the vocal tract with the objective of building an articulatory model. We were particularly interested in the consequences of stopping phonation during the acquisition. Indeed, subjects are traditionally instructed to maintain the same articulation, i.e. keeping all articulators as motionless as possible even if they are obliged to stop phonation during the acquisition. We thus exploited our system to measure the temporal evolution of the tongue to palate distance in several conditions of phonation; more details can be found in (Laprie *et al.*, 2014). Several vowels were recorded with this protocol. Here we focus on the technical aspects related to the exploitation of our acquisition system and we thus present results for one vowel only.

The tongue contour is visible in US images and the

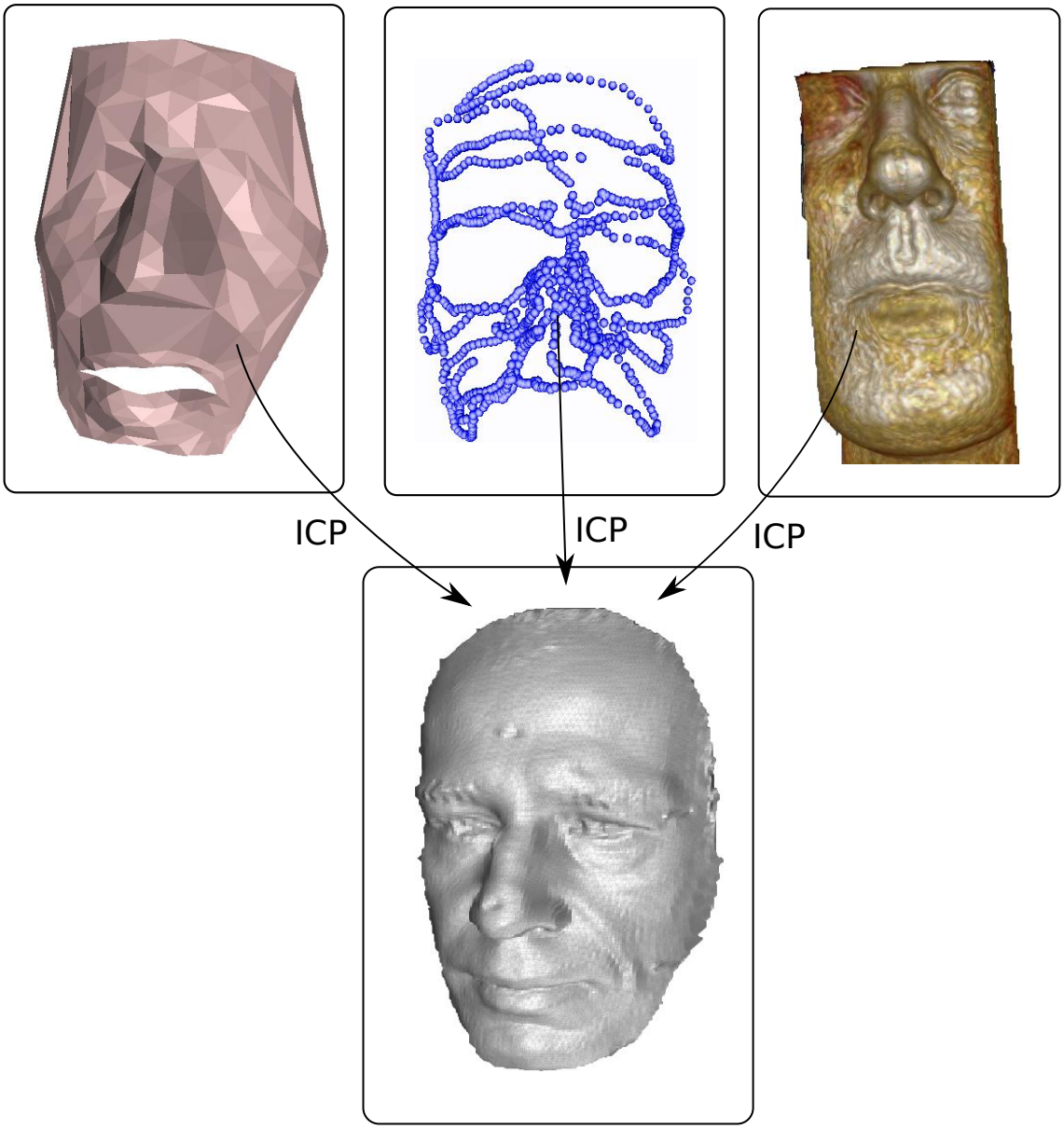


FIG. 12: Clouds of points obtained by stereo vision (left, the underlying mesh is shown), by electromagnetography (middle, by scanning the speaker's face with an electromagnetographic pointer) and by MRI (right). The digitized face (bottom) is used as the reference surface to perform ICP.

palate surface in MRI images recorded beforehand for the subject involved in this experiment. Both modalities have been merged as explained above. Concerning the impact of acquisition strategies, our system presents the strong advantage of not requiring any sensor to be glued onto the tongue like EMA. This guarantees minimal perturbation of the tongue movements (Katz *et al.*, 2006).

The US sampling frequency was set to 66 Hz and that of the Wave system to 100 Hz. Each acquisition produced an US sequence of 975 images. The tongue contour was delineated by hand every 5 images, and every 2 images when the tongue movement was fast. Two measures were realized: the minimal distance between the tongue and palate, and the place where it is minimal (see Fig. 14). The palate surface is derived from the series of sagittal slices corresponding to one MRI acquisition. The place where the distance is minimal is measured as the length of the IP curve (see Fig. 14). Point I is located at the contact point between the central incisors and the palate.

Two acquisition strategies were compared. In the first strategy the speaker was asked to stop phonation, and in the second to silently articulate the vowel. Fig. 15 shows the two articulation strategies for the vowel /u/ and enables their comparison. Fig. 16 shows on the same graphics the minimal tongue to palate distance and the place where it is minimal, for the strategy when phonation is stopped. The tongue to palate distance increases approximately 2.5 millimeters when phonation stops and

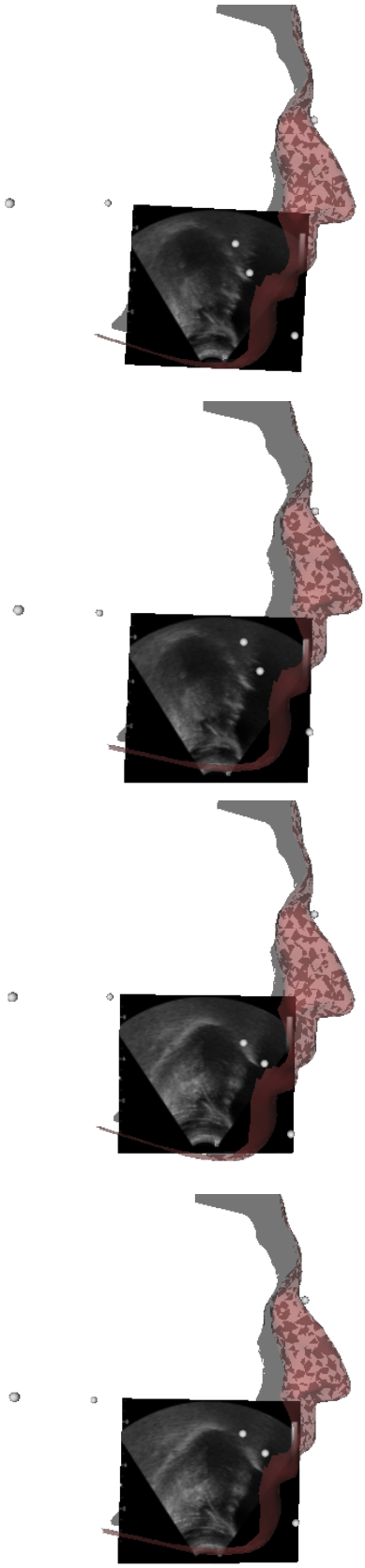


FIG. 13: Comparison of the registration by tracking the US probe (left colum) and by keeping the US probe position when acquisition starts (right column). Each row corresponds to one frame of the acquisition. Each image shows the speaker's face, the US image and the EM sensors are represented as gray spheres. There are two sensors glued behind ears, one on the nose edge, one on the US probe and two on the tongue.

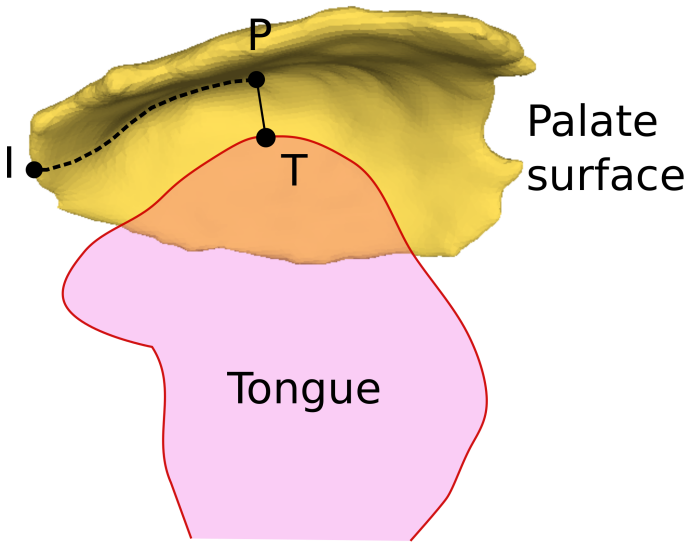


FIG. 14: Tongue to palate distance measurement. The minimal tongue to palate distance corresponds to the segment PT and the location where the distance is minimal is measured as the length of the IP curve.

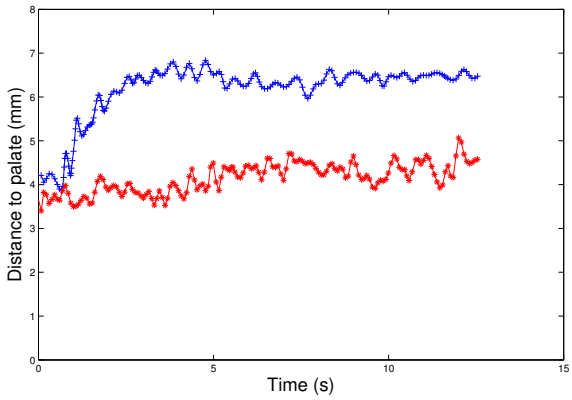


FIG. 15: Minimal distance between the tongue and the palate for the vowel /u/ in two conditions: (+) phonation stopped when recording starts, (*) silent articulation.

remains almost constant for silent articulation. Beyond the interpretation of these results, this figure shows that fine changes in the position of the tongue can be detected reliably. These measures would not have been possible without the calibration of acquisition modalities involved in the ARTIS system and merging with another modality, here MRI.

VIII. CONCLUSION

A protocol and methodology for designing a multimodal acquisition system were presented. Calibration

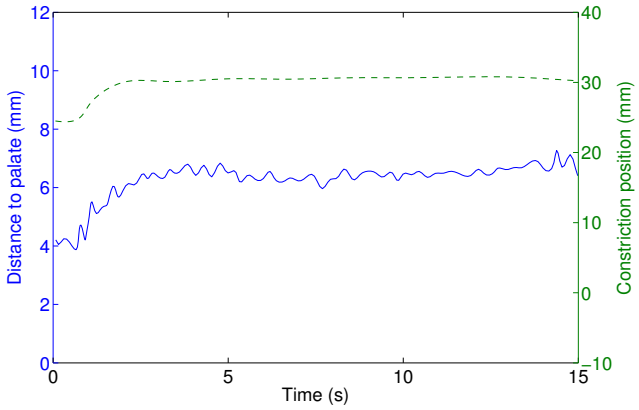


FIG. 16: Minimal distance between the tongue and the palate (solid line) and position of the palate point minimizing the distance to palate (dashed line) when phonation stops. For the second curve zero corresponds to the point I on Fig. 14.

and registration techniques used in the ARTIS system present the strong advantage of resting on techniques well founded from geometrical and mathematical points of view. Each stage of the acquisition process, either geometrical calibration or synchronization, gave rise to a rigorous assessment so as to evaluate its precision. This substantially improves the relevancy of data collected.

Furthermore an automatic procedure for fusion enables several modalities that cannot be used at the same time to be combined, for instance MRI to measure the hard palate and other walls of the vocal tract and US imaging to measure tongue movements. This is particularly interesting to investigate how the tongue form constrictions in the vocal tract and how these constrictions change over time. Finally, unlike others, our system does not require immobilization, which contributes to keep a natural articulation during acquisition of data.

Baer, T., Gore, J. C., Gracco, L. C., and Nye, P. W. (**1991**). "Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels", Journal of the Acoustical Society of America **90**, 799–828.

Besl, P. and McKay, N. (**1992**). "A method for registration of 3-d shapes", IEEE Trans. on Pattern Analysis and Machine Intelligence **14**, 239–256.

Bresch, E., Kim, Y.-C., Byrd, K. N. D., and Narayanan, S. (**2008**). "Seeing Speech: Capturing Vocal Tract Shaping Using Real-Time Magnetic Resonance Imaging", IEEE Signal Processing Magazine **May**, 123–132.

Brunner, D. O., Zanche, N. D., Fröhlich, J., Paska, J., and Pruessmann, K. P. (**2009**). "Travelling-wave nuclear magnetic resonance", Nature **457**, 994–998.

Flannery, B., Teukolsky, S., and Vetterling, W. (**1992**). *Numerical Recipes, 2nd Edition (Chapter 10)* (Cambridge University Press).

Hueber, T., Chollet, G., Denby, B., and Stone, M. (**2008**). "Acquisition of ultrasound, video and acoustic speech data for a silent-speech interface application", in *Proc. International Seminar on Speech Production*, 365–369 (Strasbourg, France).

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

Katz, W. F., Bharadwaj, S. V., and Stettler, M. P. (**2006**). "Influences of electromagnetic articulography sensors on speech produced by healthy adults and individuals with aphasia and apraxia", Journal of Speech, Language and Hearing Research **49**, 645–659.

Khamene, A. and Sauer, F. (**2005**). "A novel phantomless spatial and temporal ultrasound calibration method", in *MICCAI 2005*, 65–72.

Kirsch, S. (**2005**). "Accuracy assessment of the electromagnetic tracking system aurora", Technical Report, NDI Europe GmbH.

Laprie, Y., Aron, M., Berger, M.-O., and Wrobel-Dautcourt, B. (**2014**). "Studying MRI acquisition protocols of sustained sounds with a multimodal acquisition system", in *10th International Seminar on Speech Production (ISSP)* (Köln, Allemagne), URL `http://hal.inria.fr/hal-01002121`.

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

Munhall, K. G., Vatikiotis-Bateson, E., and Tokhura, Y. (**1995**). "X-ray film database for speech research", Journal of the Acoustical Society of America **98**, 1222–1224.

Perkell, J. S., Cohen, M. H., Svirsky, M. A., Matthies, M. L., Garabieta, I., and Jackson, M. T. T. (**1992**). "Electromagnetic midsagittal articulometer (emma) systems for transducing speech articulatory movements", Journal of the Acoustical Society of America **92**, 3078–3096.

Qin, C. and Carreira-Perpiñán, M. (**2007**). "A comparison of acoustic features for articulatory inversion", in *Proc. EUROSPEECH*, 2469–2472 (Antwerp).

Scobbie, J., Wrench, A., and van der Linden, M. (**2008**). "Head probe stabilisation in ultrasound tongue imaging using a headset to permit natural head movement", in *Proc. International Seminar on Speech Production*, 373–376 (Strasbourg, France).

Sock, R., Hirsch, F., Laprie, Y., Perrier, P., Vaxelaire, B., Brock, G., Bouarourou, F., Fauth, C., Hecker, V., Ma, L., Busset, J., and Sturm, J. (**2011**). "DOCVACIM an X-ray database and tools for the study of coarticulation, inversion and evaluation of physical models", in *The Ninth International Seminar on Speech Production - ISSP'11*, 41–48 (Canada, Montreal).

Stone, M. (**2005**). "A guide to analyzing tongue motion from ultrasound images", Clinical Linguistics and Phonetics **19**, 455–502.

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

Stone, M., Sonies, B., Shawker, T., Weiss, G., and Nadel, L. (**1983**). "Analysis of real-time ultrasound images of tongue configuration using a grid-digitizing system", Journal of Phonetics **11**, 207–218.

Story, B. H., Titze, I. R., and Hoffman, E. A. (**1996**). "Vocal tract area functions from magnetic resonance imaging", Journal of the Acoustical Society of America **100**, 537–553.

Westbury, J. R., Turner, G., and Dembowski, J. (**1994**). "X-ray microbeam speech production database user's handbook version 1.0 (139 pages)", Technical Report, Waisman Center on Mental Retardation & Human Development, University of Wisconsin, Madison.

Whalen, D., Tiede, M., Ostry, D., Lehnert-LeHouillier, H., vatikiotis Bateson, E., and Hailey, D. (**2005**). "The hask-

ins optically corrected ultrasound system (hocus)", Journal of Speech, Language and Hearing Research **48**, 543–553.
Wrobel-Dautcourt, B., Berger, M. O., Potard, B., Laprie, Y., and Ouni, S. (**2005**). "A low cost stereovision based system for acquisition of visible articulatory data", in *Proceedings of International Conference on Auditory-Visual Speech Processing (AVSP'05)*, 145–150 (Vancouver).
Zierdt, A., Hoole, P., H.G., and Tillmann (**1999**). "Development of a system for three-dimensional fleshpoint measurement of speech movements", in *Proceedings of the International Congress on Phonetic Sciences*, 73–76.