Studying MRI acquisition protocols of sustained sounds with a multimodal acquisition system

Yves Laprie[1], Michaël Aron[2], Marie-Odile Berger[1], Brigitte Wrobel-Dautcourt[1]

[1]*INRIA/CNRS/UL, LORIA, Nancy, France*
[2] *ISEN, Brest, France*
Yves.Laprie@loria.fr

Abstract

The acquisition of dynamic articulatory data is crucial to improve our understanding of speech production. Ultrasound (US) imaging presents the interest of offering a good temporal resolution without any health hazard and at a reasonable price. However, it cannot be used alone because there is no reference coordinate system and no spatial calibration. We describe a multimodal acquisition system which uses electromagnetography sensors to locate the US probe, and the method used to calibrate the US modality.
We experimented this system to investigate the most appropriate acquisition protocol for Magnetic Resonance Imaging. Three strategies were explored for one speaker: (i) stopping phonation just before acquisition, (ii) reiterated acquisitions, and (iii) silent acquisition. The measurements of the minimal tongue to palate distance show that silent acquisition generally offers the minimal articulatory drift while guaranteeing a small tongue to palate distance, i.e. a clear articulation.

Keywords: articulatory data, MRI, ultrasound imaging, multimodality

1. Introduction

Technical advances in acquiring articulatory data have often conditioned scientific breakthroughs in speech production modeling. Very substantial advances have been achieved in the acquisition of static articulatory data. By offering a millimetric accuracy 3D MRI images of the vocal tract have enabled more accurate evaluations of vocal tract acoustic modeling (Valdés Vargas, Badin, and Lamalle 2012).

On the other hand, the acquisition of dynamic geometric articulatory data still represents a challenge. Despite its recent emergence, real time MRI (Narayanan et al. 2004; Bresch et al. 2008) is not widely available because of its cost. Additionally, spatial resolution of images is still rather poor and the recording conditions (noise of the machine and supine position) alter speech production. Beside its innocuousness, the advantage of electromagnetography (EMA) is to offer a sufficiently high sampling frequency (400 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, the minimal distance to respect between two sensors to avoid aberrant measures due to magnetic interferences, and the risk of perturbing articulation with wires connecting sensors to the articulograph. On the other hand ultrasound 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. Data are recorded in the coordinate system attached to the probe.

In order to register tongue contours in a reference coordinate system it is necessary to track the position of the ultrasound probe to get its position during the acquisition. This solution has been chosen by the designers of the HOCUS system (Whalen et al. 2005). Infrared sensors are fixed onto the ultrasound 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 Nothern Digital Optotrack system. Actually, fleshpoints behind the ears are probably less mobile but the nature of the Optotrack system which utilizes infrared emitting diodes (IREDs) requires that the sensors to be visible from the cameras. The designers of the HOCUS system preferred not to add probe immobilization device so as to avoid the apparition of spurious articulatory compensation gestures. There is thus no guarantee that the probe remains in the mediosagittal plane. This setup gives the position and the orientation of the probe and head in the optical coordinate system. However, the plane of the ultrasound image cannot be known and designers thus added three sensors onto the US probe so that the plane defined by these points approximately coincides with that of the ultrasound image.

There is thus no geometric calibration since the location of a point in the ultrasound image cannot be calculated in the optical coordinate system. The geometrical transformation between sensors glued on the probe and the ultrasound image is only estimated by hand with the involved inaccuracy. Beyond this inaccuracy, it is not possible to know the position of one point of the ultrasound image in this plane since the distance between the probe and this point is unknown even if the resolution of the ultrasound machine is provided by the manufacturer.

It is thus impossible to merge US images acquired with the HOCUS system with images of the vocal tract acquired with another acquisition modality, MRI for instance. We thus developed a multimodal system (Aron, Berger, and Kerrien 2010), which combines ultrasound imaging and electromagnetography. Ultrasound imaging allows tongue tracking while electromagnetography is used to track the position of the ultrasound probe and enables the registration of ultrasound with other modalities, here MRI for instance. One of the main contributions is the spatial calibration of the imaging modalities, especially ultrasound, with respect to the modality used to merge all data, here electromagnetography.

The calibration of an imaging modality with respect to another is generally obtained by considering points visible in both modalities. In our case, the direct calibration of both modalities is impossible because electromagnetic sensors are invisible in ultrasound images. It is thus necessary to design an object,

called phantom, whose geometrical properties are known and which is 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 easiness of use, accuracy, and precision. There is no agreement about the best phantom design. The phantom we designed is inspired from that of (Khamene and Sauer 2005). Two 5DOF sensor coils were fixed at both extremities P_0 and P_1 of a rigid wood stick approximately 25 cm long and 3 mm on diameter (as seen on Figure 1).

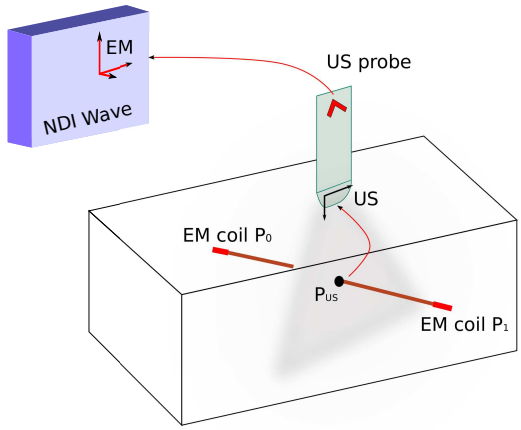

Figure 1: *Experimental setup used for ultrasound calibration.*

This line segment whose equation is known from the electromagnetic sensors fixed at extremities is easily detectable in the US images. The line pointer and the transducer were immersed into water at the room temperature, and thirty images corresponding to different positions and orientations of the transducer were acquired. 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. The uncertainty due to noise was overcome 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 ($\approx 1540m/s$). Bilaniuk (Bilaniuk and Wong 1993) showed that the speed of sound in water at 20° C is $1485m/s$. Therefore every point in US images was corrected by shortening its depth (distance to the US focal point) by a ratio of $1540/1485 \approx 1.04$.

Since the first developments (Aron, Berger, and Kerrien 2010) the system has been substantially modified: (i) it utilizes the new NDI Wave system which offers a better sampling frequency, 100 Hz instead of 50 Hz in the previous system, (ii) it also offers a better calibration of the ultrasound modality and (iii) provides a simpler and, therefore more robust, synchronization system which requires only one PC used to supervise the acquisition system. One of the advantages provided by the calibration of the ultrasound modality is to enabling the fusion with MRI data since the 3D position of points in the ultrasound images is expressed in the electromagnetic coordinate system. MRI images are recorded beforehand and are fused with ultrasound data. This enables the calculation of the minimal distance between the tongue and the palate.

2. Experiment to asses MRI acquisition protocols

We report here first acquisition experiments aiming at studying protocols used to acquire static 3D MRI images of the vocal tract. We were particularly interested in the consequences of stopping phonation during the acquisition. Indeed, subjects are traditionally asked 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. Other solutions consist of stopping phonation just before the acquisition starts, repeating several short acquisitions synchronized with acquisition by identifying voice activity or even using silent speech.

A good acquisition protocol should guarantee a steady position of the tongue to get good quality MRI images together with relevant positions of the tongue. We evaluated the tongue position by measuring the minimal distance between the tongue and the palate. Indeed, we suspected that the tongue lowers as soon as phonation stops, and this distance gives interesting information about the vowel quality because too big a distance corresponds to a centralization.

The tongue contour is visible in ultrasound images and palate surface in MRI images. 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 traditional EMA which is likely to alter articulation (Katz, Bharadwaj, and Stettler 2006). Here EMA is only used to track the US probe and to cancel head movements. This guarantees minimal perturbation of the tongue movements.

We investigated three conditions:

1. The speaker starts phonating the sound and then stops after 5 seconds while maintaining the same articulation for approximately 10 seconds,
2. The speakers phonates the vowel for 3 seconds approximately, stops phonation and starts again phonation. He reiterates this scheme during the 15 seconds of the acquisition and is allowed to take breath.
3. The speaker silently articulates the vowel for 15 seconds.

The four vowels /u,i,a,y/ have been recorded with this protocol. The sampling frequency of ultrasound was set to 65Hz and that of the wave system to 100 Hz. Each acquisition produced an ultrasound film of 975 images. The tongue contour was delineated by hand every 5 images, and every 2 images when the tongue movement is fast. Due to the lack of space, we only inserted figures for /u, i, a/. Figures 2, 3, 4 show the the three articulation strategies for each vowel. They enable the comparison of the strategies from the stability point of view. The three vowels exhibit the same tendency but with marked differences in terms of relative amplitudes. The strategy of stopping phonation always gives rise to an articulatory drift, i.e. an abduction movement of the tongue, with a marked fast transition for /u/ and almost a linear transition for /i/. For these two vowels the distance to palate increases between 3 and 4 mm. On the other hand, the drift is smaller for /a/, about 1 mm, but the initial distance was also bigger at 8mm and the main articulatory characteristic is the narrow tube in the pharyngeal region.

At the other extremity, the strategy of silent speech gives rise to a limited drift within approximately a millimeter, and slightly more for /u/. Indeed, unlike /i/ and /a/ for which the tongue can be maintained in the same posture because it can contact molars (/i/) or the mouth floor (/a/) /u/ requires a more

substantial effort without the contribution of teeth or mouth floor to adjust the posture. This is likely to explain the slight abduction movement observed for silent /u/. The reiterating phonation gives rise to strong oscillations of the tongue in both directions, i.e. very low minimal distances between tongue and and palate, but also large values of this distance when phonation stops. This tendency is not as marked for /a/ probably because mouth opening necessary to produce /a/ corresponds less to the movement of the tongue than that of the mandible, which has a greater inertia. Depending on the vowel the amplitudes of oscillations of the reiterating strategy are well between these two extremities for /a/, reach the values of the stopping and silent strategy for /u/ and exceed them in the case of /i/. In the latter case it should be noted that the articulatory drift accompanying the stopping strategy is not complete at the end of the recording. It is likely that the maximal value of the oscillations is the limit of this drift.

We also evaluated the horizontal movement of the constriction during acquisition. A remark should be made about this measure. The horizontal position of the highest point of the tongue contour drawn on ultrasound images is not very precise because the constriction of a vowel is not very strong and therefore extends over a fairly large part of the palate and also because the highest point is generally located in a flat part of the tongue contour. We thus only give this measure for the stopping strategy. As shown by Figures 5, 6, 7 the constriction location moves backwards approximately 5 mm for /u/. This movement coincides with the abduction tongue movement.

This experiment was intended to choose the best strategy for acquiring MRI image for one specific subject. Before this experiment we expected that the reiterating strategy could favor the realization of articulatory targets close to those of sustained phonation. To some extent this strategy avoids the abduction movement of the tongue, with the counterpart that it is difficult to realize the same target several times without the help of an imposed phonetic context. Besides, the amplitude of the articulatory effort necessary to restart phonation is another explanation and could trigger a strong movement of the tongue upwards which blurs the results.

The silent speech strategy turns out to give a good stability of the tongue even if there is no guarantee that the articulatory configuration corresponds to that of the target vowel. However, it can be seen that the minimal distance between the tongue and palate is close to that of phonation. In addition, the horizontal movement of the maximal constriction seems to be sufficiently small to ensure that the tongue shape is reasonably close to that of the phonated vowel. This led us to keep the silent strategy with prior phonetic training to guarantee the relevancy of vocal tract shapes recorded with MR imaging.

3. Conclusion

This first experiment with the new version of our multimodal system shows that acquisition protocols used in MRI should not involve changes in phonation, i.e. stopping phonation when the acquisition lasts too long, or reiterated articulatory gestures. We will now test other speakers in order to compare different strategies that can be used for MRI acquisitions of static vocal tract shapes. Three-dimensional data require an acquisition duration of approximately 15 seconds. By taking into account the additional time required for the subject to reach the target before the noise of the MRI machine starts it is difficult to sustain phonation up to the end of the recording. It seems that creaky voice, as proposed by A. Baker `http://www.ultrax-speech.org/news/mri-vocal-tract-data-collaboration`, could allow the speaker to sustain phonation longer while keeping articulatory targets very close to those of natural speech. We will include this strategy in future experiments.

To our knowledge this is the first time that MRI acquisition protocols are studied from the point of view of the of image stability they offer without perturbing speech production. Indeed, this acquisition system imposes no constraint on the tongue movement and does not alter the environment with the presence of bulky devices. Up to now we are using a pediatric probe held by one assistant, who checks the orientation of the probe via a graphical interface. Indeed, we preferred this solution to an helmet which could contain some ferromagnetic metal and also limits the jaw opening (Jaumard-Hakoun et al. 2013). Since the weight of the probe is sufficiently low relative to the mandible, another solution would consist in fixing the probe on the subject's under the jaw with a self-adhesive stretchable strip that easily adapts to jaw and neck contours.

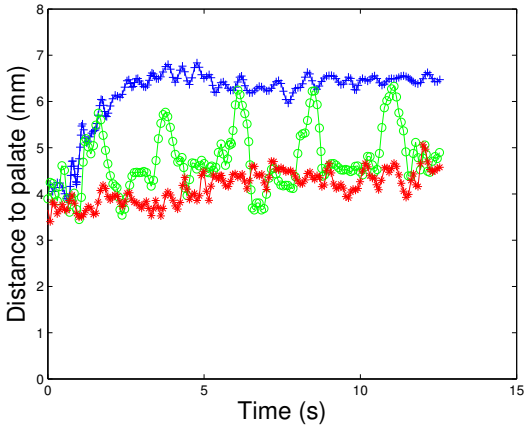

Figure 2: *Production strategies for /u/. In blue phonation stopped when recording starts, in red silent articulation and in green reiterated phonation*

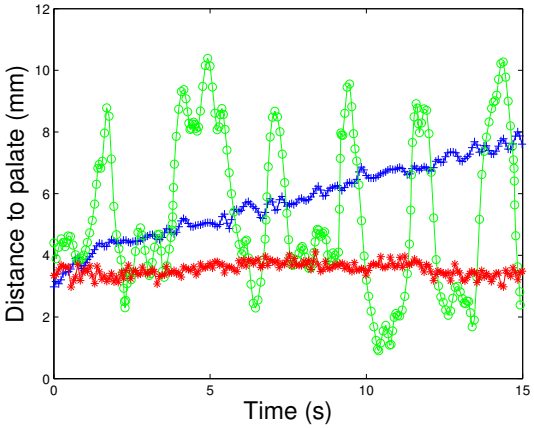

Figure 3: *Production strategies for /i/. In blue phonation stopped when recording starts, in red silent articulation and in green reiterated phonation*

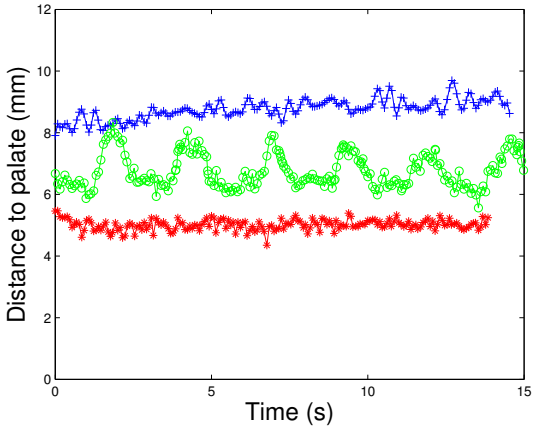

Figure 4: *Production strategies for /a/. In blue phonation stopped when recording starts, in red silent articulation and in green reiterated phonation*

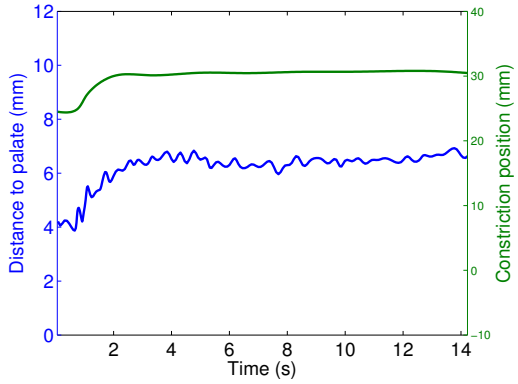

Figure 5: *Distance to palate (in blue) and location of the constriction (in green) for /u/. The location of the constriction is given from the first point of the palate located behind the point where upper incisor and palate connect.*

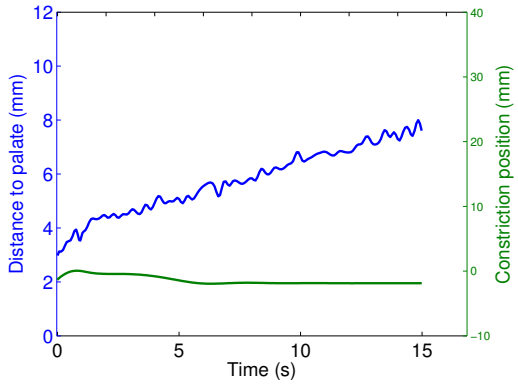

Figure 6: *Distance to palate (in blue) and location of the constriction (in green) for /i/. The location of the constriction is given from the first point of the palate located behind the point where upper incisor and palate connect.*

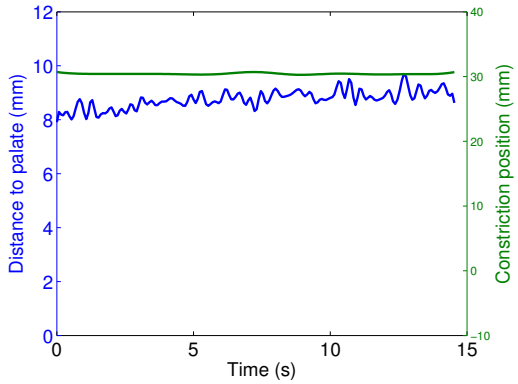

Figure 7: *Distance to palate (in blue) and location of the constriction (in green) for /a/. The location of the constriction is given from the first point of the palate located behind the point where upper incisor and palate connect.*

4. References

Aron, M., M. O. Berger, and E. Kerrien (2010). "Evaluation of the uncertainty of multimodal articulatory data". In: *Ultrafest V*. New Haven, USA.

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

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

Jaumard-Hakoun, A., S. K. Al Kork, M. Adda-Decker, L. Amelot A. andBuchman, T. Fux, C. Pillot, P. Roussel, M. Stone, G. Dreyfus, and B. Denby (2013). "Capturing, Analyzing, and Transmitting Intangible Cultural Heritage with the i-Treasures Project". In: *Ultrafest VI*. Edinburgh, UK.

Katz, W., V. Bharadwaj, and P. Stettler (2006). "Influences of Electromagnetic Articulography Sensors on Speech Produced by Healthy Adults and Individuals With Aphasia and Apraxia". In: *Journal of Speech, Language and Hearing Research* 49, pp. 645–659.

Khamene, A. and F. Sauer (2005). "A novel phantom-less spatial and temporal ultrasound calibration method". In: *MICCAI 2005*, pp. 65–72.

Mercier, L., T. Lango, F. Lindseth, and D.L. Collins (2005). "A review of calibration techniques for freehand 3-D ultrasound systems". In: *Ultrasound in Med. and Biol.* 31.4, pp. 449–471.

Narayanan, S., K. Nayak, S. Lee, A. Sethy, and D. Byrd (2004). "An approach to real-time magnetic resonance imaging for speech production". In: *Journal of the Acoustical Society of America* 115.4, pp. 1771–1776.

Valdés Vargas, J. A., P. Badin, and L. Lamalle (2012). "Articulatory speaker normalisation based on MRI-data using three-way linear decomposition methods". In: *Interspeech 2012 (13th Annual Conference of the International Speech Communication Association)*. Portland, Oregon, USA.

Whalen, D. H., K. Iskarous, M. K. Tiede, D. J. Ostry, H. Lehnert-Lehouillier, E. Vatikiotis-Bateson, and D. S. Hailey (2005). "The Haskins optically corrected ultrasound system (HOCUS)". In: *Journal of Speech, Language, and Hearing Research* 48.3, pp. 543–553.