Multimodal Fusion of Electromagnetic, Ultrasound and MRI Data for Building an Articulatory Model

Michael Aron, Marie-Odile Berger, Erwan Kerrien

LORIA/ INRIA Nancy Grand-Est, BP 101, 54602 Villers les Nancy, France
E-mail: {aron,berger,kerrien}@loria.fr

Abstract

Data fusion from multiple sensors is of significant interest to the speech research community, as it can potentially provide a better picture of speech production through the use of complementary sensor modalities. This paper deals with the practical aspects of this problem, such as acquisition and processing of the dynamic ultrasound (US) and electromagnetic (EM) data of the tongue during speech production, static MRI images of the vocal tract using repetitions, and registration of the data from these different sources to a common reference frame. To the best of our knowledge, this is the first work that demonstrates the potential of static and dynamic data fusion in the construction of articulatory databases.

1 Introduction

A major goal of the speech research is to be able to build a dynamic model of the vocal tract of a speaker. An ideal imaging system should cover the whole vocal tract (from larynx to lips) and the face, have a sufficient spatial and time resolution, and not involve any health hazard. To date, no single modality satisfies all these requirements. Stone [5] uses US images to get high frequency images of the tongue shape in the mid-sagittal plane of the head, Badin [3] uses static MRI images which covers in 3D the whole vocal tract at a good spatial resolution but for sustained sounds only.

Therefore, data fusion is not only a viable, but also an essential strategy to build a dynamic and a 3D model of the vocal tract. In this paper, a framework to acquire, register and process data together is presented. The proposed system utilizes 3D magnetic resonance imaging (MRI) for sustained sounds, ultrasound (US) imaging to get images of the tongue with a high frame rate, and electromagnetic (EM) sensors to complete the information on the apex. Besides, registration techniques to combine these several data sources are proposed in order to fuse the information provided by each modality.

2 Dynamic data: US and EM

2.1 Acquisitions

In our work [2], a system to get temporally and spatially aligned acquisitions of US and EM data of the tongue was presented. EM data are used to get the position of the apex of the tongue which is not always visible on the US images.

A corpus of US images with EM data was acquired for a speaker, including Vowel-Vowel (/ae/...), Vowel-Consonant-Vowel (/aka/...) and 120 complete sentences in French. The length of this corpus is 615 seconds (40 590 US images).

2.2 Extracting the tongue contours

In order to be able to process the large amount of data in the corpus efficiently, an automatic tool to track the tongue shape on the US images (i.e. the mid-sagittal plane of the head) has been developed [2]. This tool combines a preprocessing of US images, the estimation of the displacement via optical flow, snake with constraints on the extremities, and use of EM sensors as prior to help the tracking. Our method was evaluated via a comparison with manual tracking (considered as the ground truth position of the tongue shape) and the tracking tool EdgeTrak [5]. Results, presented in table 1, illustrate that our method outperforms EdgeTrak in two areas: tracking is more robust in terms of precision, and less expensive in terms of computational cost. Even if, as in Edgetrak, human intervention is still necessary in order to help tracking on highly blurred US images, the time needed for correction has significantly decreased.
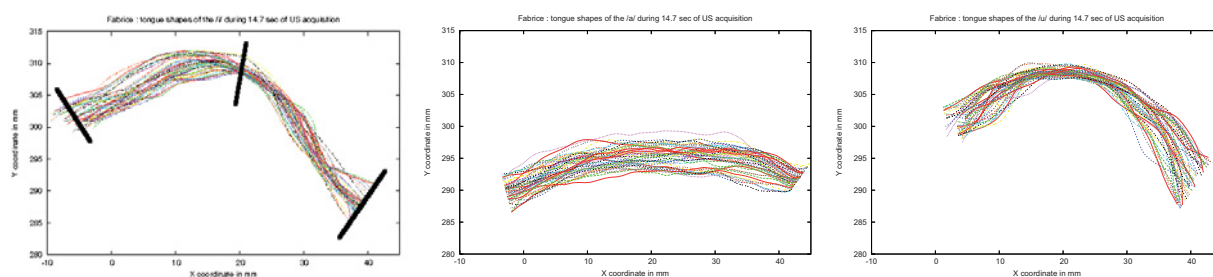
Figure 1: *Into the EM coordinate system attached to the US probe, tongue shapes variability during 14.7 seconds of US acquisition. Apex is on the left-side of the images, back of the tongue on the right-side. Left: /i/. Middle: /a/ Right: /u/.*

Table 1: *Results of the tracking on the group of phonemes /ao/ and /au/ (190 images - 2.9 sec).*

	Mean error [Std. dev] (mm)	% images err>2mm	Process time (min)
EdgeTrak	5.68 [2.57]	93.2	10
Our tool	0.97 [0.34]	1.6	2

3   Static data: MRI

3.1   Existing work

Many 3D models of the vocal tract reconstructed from MRI data have been proposed over the years. With the exception of a few studies employing experimental dynamic MRI techniques (Shadle [4]), the majority of studies still employ static MRI, which does not require any specific and expensive hardware. The major difficulty encountered when using MRI for speech analysis is its quite long acquisition time. As a result, it can only be used for sustained sounds.

Badin [3] acquired 53 sagittal slices of the head. The slices were sampled every 4.0 mm, with a thickness of 3.6 mm, in images of 128x128 pixels. They were acquired in 43 sec. Therefore the subject had to artificially sustain the articulation, breathing out very slowly or in full apnea if possible.

In [6], an acquisition protocol with pauses has been proposed. The slice thickness was 5 mm, with a resolution of 0.938 mm/pixel in 26 sagittal images of 256x256 pixels. Making pauses during the acquisition allowed the sound to be actually uttered during time spans of 8 seconds, enabling the speaker to take a breath between two sound utterances: images were thus acquired in 266 seconds.

The first protocol has the advantage to be shorter in time but it requires to artificially sustain a sound.

The question arised whether a speaker is able to maintain the articulators at a correct position if not actually uttering the sound.

3.2   Variability of the tongue position

A simple experiment was conducted with our US system: a speaker was asked to maintain his tongue at the same position, by breathing smoothly, during 43 sec for various phonemes. Between the start and the end of the acquisition, the spreading of the positions on a point at the middle of the tongue was superior to 15 mm for phonemes with a highly curved tongue (such as /i/ or /u/).

This variation must be compared with the natural motion of the tongue observed when the speaker is actually uttering the sound. For three vowels /i/, /a/, /u/ which can be sustained, a 15 sec US sequence was acquired. Tongue shapes were extracted with our tracking tool (Fig.1).

Table 2: *Variability in mm of the US tongue position for 3 phonemes in an US acquisition.*

	apex	middle	back	mean
/i/	6	3	6	5
/a/	7	7	6	7
/u/	6	3	9	6

The amplitude of the position variations was measured on 3 sections of the tongue (Fig.1,left). The US probe movements were removed. The average of this amplitude is 6 mm (Table 2). Moreover, on /i/ and /u/, the apex and the back of the tongue have greater amplitude than the middle of the tongue. This can be explained by the place of constriction of vowels which has a stabilizing effect for the tongue shape positions.

According to these experimentations, the strategy

of maintaining artificially the position of the articulators without emitting the sound induces abnormally high variations on the tongue position. Therefore, the MRI protocol with pauses was chosen to get acquisitions of the vocal tract with the best quality.

3.3 Acquisition setup

The MRI machine (1.5T, GE Healthcare) at Nancy Hospital was used in 2D Spin Echo mode to get 32 sagittal slices. The slice thickness was 3 mm, sampled every 2.5 mm, in images of 512x512 pixels. The acquisition time was 18 seconds, repeated 6 times. For one sound, the acquisition time was 108 sec.

There are numerous advantages in using such a protocol: blurred images are avoided because of the short acquisition time, and compared to previous works, the image quality is improved with larger image size and a better resolution (a voxel is $0.625 \times 0.625 \times 2.5$ mm). Finally, the whole head is covered, making spatial registrations possible between different acquisitions.

A high-resolution MRI of the speaker's head was also acquired in order to have a detailed reference MRI scan of the speaker's head. This detailed scan is used as the reference coordinate system in the next section to set into spatial correspondence all the different acquisitions obtained with our protocol.

4 Multimodal registration

4.1 Method

Fusing US and MRI data requires to register both modalities. This registration consists in estimating the rigid transformation (rotation + translation) T_{us2mri} between the MRI and the US coordinate systems. The EM coordinate system is used as an intermediate to compute this rigid transformation.

Firstly, the rigid transformation $T_{us2emprobe}$ between the US coordinate system and the EM coordinate system attached to the US probe is computed via the calibration procedure explained in [1].

Then, the rigid transformation $T_{emprobe2emhead}$ between the coordinate system attached to the US probe and the EM coordinate system attached to the speaker's head is estimated.

Finally, the rigid transformation $T_{emhead2mri}$ between the EM coordinate system attached to the speaker's head and the MRI coordinate system is calculated thanks to an Iterative Closest Point method: the searched transformation T is the one which minimizes the distance between the points P_i of the face

pointed with the EM system and the reconstructed surface S of the head from MRI:

$$T_{emhead2mri} = \min_T \sum_i dist(T(P_i), S)$$

The composition of these three rigid transformation gives the complete transformation T_{us2mri} between the US modality and the MRI modality (Fig.2).

4.2 Results

Once registration is performed, MRI and US can be fused.

On Fig.3, an extracted shape of the tongue and the palate from an MRI are displayed on a US image. Because of the variability of the tongue shape (Sec.3.2), the US image has been chosen to correspond with the superimposed MRI tongue curve.
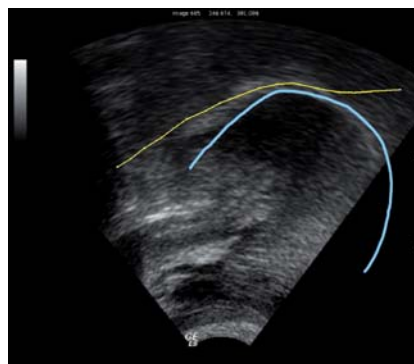

Figure 3: *MRI tongue shape (in blue, thin line) and palate (in yellow, thick line) extracted from /u/ on MRI and displayed on a US image.*
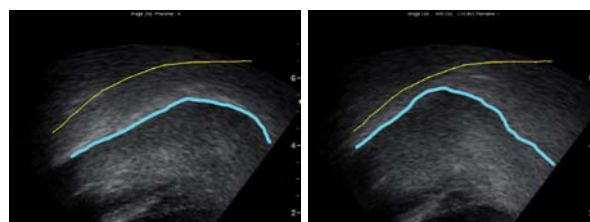

Figure 4: *MRI palate (in yellow, thin line) displayed on US images with tracked tongue shape (in blue, thick line). Left: /a/. Right: /i/.*

It is of special interest to display the palate extracted from an MRI on the dynamic US images to recover the place of the main constriction. On Fig.4, the palate extracted from the MRI is displayed on a US image, with the tongue shape extracted by our tracking tool. We are currently testing these fused
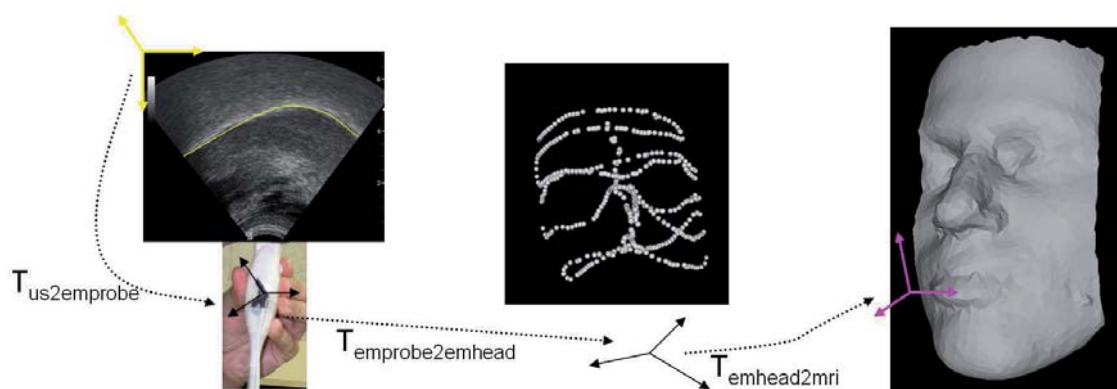
Figure 2: *Several steps are used for the calculation of the registration between the US modality and the MRI modality.*

data for the estimation of the articulatory parameters of the vocal tract built from Maeda's model.

The validation is twofold. First, recovering articulatory parameters from US images and checking that the dynamics is consistent with that observed on X-ray moving pictures. And second, resynthesizing the speech signal from vocal tract shapes derived from registered US data and comparing it with the original speech signal. These two assessment criteria ensure that vocal tract shapes are consistent with gestures realized by the speaker. This validation can be carried out on VV sequences implying large articulatory gestures, i.e. /ia/, /iu/ for instance. It can also be used for VV sequences implying more subtle gestures, /ie/ or /iy/ for instance.

5   Conclusion

This paper presents a method for extracting and fusing different sources of speech production data. An automatic tool is used for the tracking of the tongue shape in US images. A study of variability of this position has been quantitatively performed. An high resolution MRI protocol using pauses has been discussed in order to limit the variations of the tongue shape positions for sustained sounds. Finally, a registration method between MRI and US has been presented, by displaying the positions of the tongue and the palate extracted from MRI on US images. To our knowledge, this is the first prototype which automatically fuses MRI, EM and US data for the construction of an articulatory database

References

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

[2] M. Aron, A. Roussos, M.-O. Berger, E. Kerrien, and P. Maragos. Multimodality Acquisition of Articulatory Data and Processing. In *Proc. of Eusipco 2008*, Lausanne, Switzerland, 2008.

[3] P. Badin, G. Bailly, L. Revéret, M. Baciu, C. Segebarth, and C. Savariaux. Three-dimensional articulatory modeling of the tongue, lips and face, based on mri and video images. *Journal of Phonetics*, 30(3):533–553, 2002.

[4] C. Shadle, M. Mohammad, J. Carter, and P. Jackson. Multi-planar dynamic magnetic resonance imaging: new tools for speech research. In *ICPhS*, 1999.

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

[6] B. H. Story, I. R. Titze, and E. A. Hoffman. Vocal tract area functions from magnetic resonance imaging. *Journal of Acoustical Society of America*, 100(1):537–553, 1996.