Acquisition and synchronization of multimodal articulatory data

Michael Aron, Nicolas Ferveur, Erwan Kerrien, Marie-Odile Berger, Yves Laprie

Acquisition and synchronization of multimodal articulatory data

Michael Aron, Nicolas Ferveur, Erwan Kerrien,
Marie-Odile Berger, Yves Laprie

INRIA Lorraine - CNRS UMR7503 - Nancy-Université
615, rue du Jardin Botanique. 54602 Villers-lès-Nancy, France
{aron,ferveur,kerrien,berger,laprie}@loria.fr

Abstract

This paper describes a setup to synchronize data used to track speech articulators during speech production. Our method couples together an ultrasound, an electromagnetic and an audio system to record speech sequences. The coupling requires a precise temporal synchronization, to know exactly the delay between the recording start of each modality, and to know the sampling rate of each modality. A complete setup and methods for automatically synchronizing data are described. The aim is to get a fast, low-cost and easily reproducible acquisition system in order to temporally align data.

Index Terms: articulatory data, speech production, ultrasound, electromagnetic sensors, temporal synchronization.

1. Introduction

Among all the acoustic signals human beings are exposed to, speech plays a particular role because it is probably the most important and efficient modality of communications. In addition, human beings can produce and perceive speech. The impact of a better comprehension of speech production thus covers several theoretical and applied areas of automatic speech processing: links between production and perception by searching for common cognitive representations [1], automatic speech recognition by searching for the critical articulators [2], talking heads by providing better coarticulation models, acoustic-to-articulatory inversion by enabling the acquisition of data to construct an analyzing model...

Observing the speaker's vocal tract and face thus gets a crucial importance. An ideal imaging systems should cover the whole vocal tract (from larynx to lips), the face, have a sufficient spatial and time resolution, and not involve any health hazard.

Several methods have been proposed over the years to measure the vocal tract. X-ray imaging offers a good time resolution and covers the whole vocal tract but does not provide 3D images and more importantly involves health hazards. It thus has been given up even if the processing of existing cineradiographic databases [3] recorded in the seventies and eighties may provide interesting articulatory information. Even if MRI imaging does not offer a sufficient fast sampling rate it provides 3D images of the vocal tract with a good spatial resolution. It is thus now widely used to collect still images and derive 3D models [4] from these images. The Electromagnetic Midsagittal Articulograph (EMMA) enables investigation of speech articulators dynamics, an particularly the tongue [5]. However, the tongue shape is given by a very small number of points and wires connecting coils to the EMMA system slightly perturb speech articulation. Finally, ultrasound (US) imaging offers a good time resolution and a continuous 2D space resolution but only for the portion of the tongue which is not hidden by the front part of the jaw, or invisible due to the air of the sublingual cavity between the tongue tip and US probe. This means, that most of the time, and especially for front vowels and dental consonants, the tongue tip is not visible.

Since no single imaging technique answers the requirements mentioned above it is thus necessary to combine several modalities. Several attempts have been made to acquire multimodal articulatory data. The system HOCUS [6] combines ultrasound imaging together with infrared emitting diodes placed on the lips and on the probe. In the HATS system [8], M. Stone used several 2D US acquisitions to recover a 3D model of the tongue. Despite the fact that synchronizing the data is important especially for fast speech utterance, the problem of synchronizing the data in these systems is generally not addressed or manually performed: automatic synchronization of US images with sound is not possible in the HOCUS system. In [8], misalignments between audio and videos may reach 30 frames. In addition, in [8] and in [6], the fusion between the sound and the US video sequence is achieved by downsampling the video framerate at 30 fps. We thus propose in this paper an acquisition system which integrates an automatic synchronization of the articulatory data and where each modality keeps its own sampling rate. This system combines ultrasound imaging to get the tongue shape, two electromagnetic (EM) sensors glued onto the tongue (tongue tip and dorsum) to recover the tongue tip position, and stereovision cameras to get the speaker's face shape. Additional electromagnetic sensors are used to get the geometrical locations of the ultrasound probe and of the subject's head.

Combining several acquisition modalities requires that all geometrical and temporal data be consistent together. A calibration of each modality involved is thus necessary to relate data acquired to a common system geometrical coordinates. Calibration procedures have been described in our previous work [7]. In this paper, we focus on the automatic synchronization of the data. A whole setup, easily reproducible by a user, and including US, EM and audio is presented to automatically fuse all the data together.

2. System setup

2.1. General description

2.1.1. Synchronization principle

The different modalities (the EM system, the US system and the audio recording system) are supervised by a control PC. These modalities are synchronized together by using an external event, which is an audio beep emitted by the control PC. As this PC also owns an internal clock (used as the reference base time), the

audio beep signal and a signal within the modality to be synchronized can be temporally aligned using this reference base time.

2.1.2. Overview

Our setup includes an US machine (a Logiq5 Expert, General Electrics), an EM system (Aurora, Nothern Digital Inc.), and a control PC. The audio recording of the speaker's voice is made through a microphone plugged into an audio recorder (actually a PC in our setup). Another microphone is used for the synchronization of data by recording audio beeps emitted by the control PC. The setup between all the different devices used for the speech acquisition is summarized on Fig. 1.

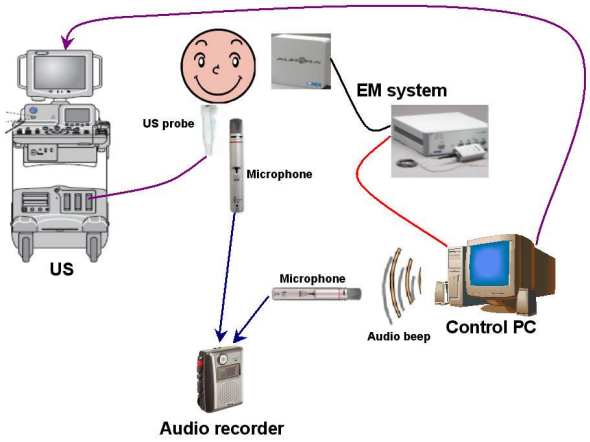


Figure 1: *The setup.*

2.2. Specificities and requirements

2.2.1. US recording process

Regarding the US video recording, images are stored in a cineloop on the US machine: all images corresponding to the last 15 seconds are recorded in a video buffer, and as soon as the user pushes a button or a footswitch, images in the buffer are saved.

2.2.2. Synchronization requirements

As the modalities have their own time scales, which generally don't agree, synchronizing several modalities requires two points:

- the knowledge of the delay between the start of acquisition of all the modalities.
- the knowledge of the sampling rate of each modality with a sufficient accuracy.

The sampling rate of each modality must be measured. Indeed the real value could be slightly different from the theoretical value given by the manufacturer. Because it is a cumulative error, little imprecision on the sampling rate yields non negligible errors after few seconds. For example, an audio recorder could have a sampling rate of 43.2 kHz instead of the theoretical value of 44.1 kHz (2% error). After 15 seconds of acquisition, it provides a shift of 300 ms (approximately 20 US images in our setup). In addition, acquisition frequency is sensitive to temperature. Therefore, estimating these values for each acquisition helps to reduce significantly the temporal shift.

2.2.3. Material characteristics

The main acquisition characteristics of each modality are summarized in Table 1.

Table 1: *Main acquisition characteristics.*

	EM	US	Sound
Acquisition rate	40 Hz	66 Hz	44100 Hz
Recording time	unlimited	15 seconds	unlimited
Data format	TXT files	DICOM	WAV files
Recording process	real time	cineloop	real time

3. Temporal synchronization

3.1. EM-Audio synchronization

3.1.1. Principle

The aim is to synchronize the EM time scale (value given by the EM system for each measure) and the audio time scale (time of each audio sample). The control PC simultaneously emits a beep and register its clock value (C_{beep}). By identifying the emitted beep on the recorded audio file (T_{beep}), the delay between the PC's time scale and the audio's time scale can be computed (eq.1).

$$\{Audio\_PC\}_{delay} = |T_{beep} - C_{beep}| \quad (1)$$

Along the same line, the control PC simultaneously registers its clock value (C_{em}) at the first EM sample request. The delay between the PC's time scale and the EM's time scale is computed using this value and the time value given by the EM system for the first sample (T_{em}) (eq.2).

$$\{EM\_PC\}_{delay} = |T_{em} - C_{em}| \quad (2)$$

The delay between the audio's time scale and the EM's time scale is computed using eq.1 and eq.2 :

$$\{EM\_Audio\}_{delay} = ||T_{beep} - C_{beep}| - |T_{em} - C_{em}|| \quad (3)$$

Fig. 2 summarizes this synchronization principle.

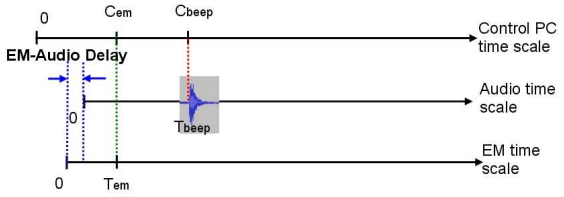


Figure 2: *Synchronization between EM and audio data.*

The synchronization accuracy depends on several parameters, the most significant being the EM acquisition period, others being negligible (time propagation of the signal, clock accuracy...)

3.1.2. Experimental validation and results

In order to prove that the accuracy of the estimated delay between EM data and audio data mainly depends on the EM acquisition period, we conducted the following experience. The principle was to use a common event visible within the two modalities. The user hit a microphone with an EM sensor. This event was visible both on the EM signal and on the audio signal. The two signals were temporally aligned and the residual delay was measured as seen on Fig. 3.

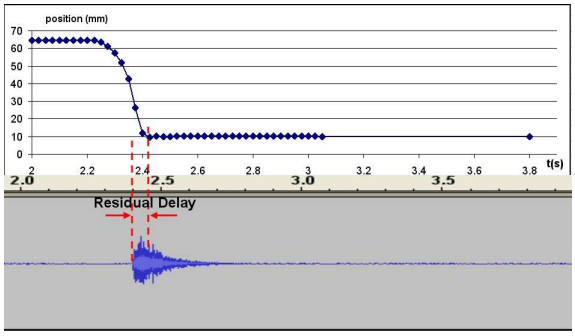


Figure 3: *Measure of the residual delay between EM data (top) and audio data (bottom).*

This experimentation was repeated several times to measure the time delay repeatability. The values found were always less than the EM acquisition period (25 ms). As a consequence, it validated our synchronization method.

3.2. US synchronization

3.2.1. Principle

The process of recording data on the US system is made a posteriori, using a cineloop. Therefore, US data are temporally synchronized with the other modalities using the US system footswitch input. This input is controlled by a relay piloted by the PC's parallel port output. A pulse sent on this output simulates a push on the footswitch and triggers the US recording. By registering the PC's clock value simultaneously with the parallel port switch, the last US frame is synchronized with the PC's time scale and therefore with the other modalities as seen on Fig. 4.

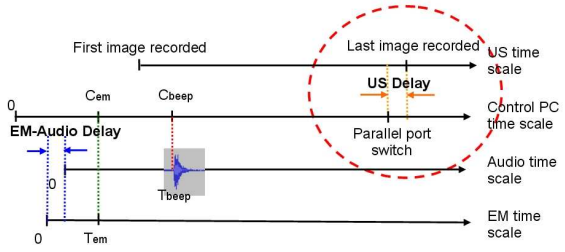


Figure 4: *Synchronization between US, EM and audio data.*

The delay between the footswitch pulse and the last recorded US image is not specified by the manufacturer. It has to be determined by an experimental protocol.

3.2.2. Experimental determination of the US delay

In order to compute the delay between the US recording process and the parallel port switch, the following experience was proposed: an EM sensor was put into water, fixed on a stick and used to hit a plastic container. This movement was visible onto US and can be correlated with the sound emitted when the sensor hits the container. This experimentation was repeated several times, to measure repeatability of the time delay. A video showing this experience can be found on our website, http://magrit.loria.fr/Confs/Interspeech07

The delay found was 52 ms with a standard deviation of 14 ms for 10 experimentations. This accuracy corresponds in our setup to an accuracy of ± 2 US images.

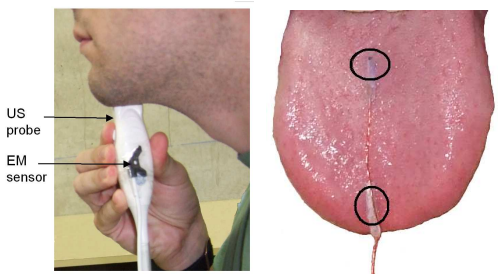
4. Experimental results

4.1. Setup

The experimentation consisted in putting an US probe under the chin (Fig. 5 a) to get the tongue contour in the midsagittal plane of the head. Two EM sensors were also glued on the tongue in order to complement the information provided by the US image (Fig. 5 b): one sensor was glued on the tip of the tongue (apex), and one other sensor was placed on the tongue dorsum.

To spatially reference EM data and US images in the same way, a sensor was mounted on the US probe to track its motion. Using this sensor, EM data provided by the sensor on the tongue can be expressed within the US system coordinate.

The system was tested on a French native speaker. The corpus included several tongue movements, including different speeds and different amplitudes: VCVs and two sentences.



(a) (b)

Figure 5: *(a):US probe under the chin with an EM sensor on it. (b): two EM sensors glued on the tongue*

4.2. Spatial calibration

In [7], we have proposed a calibration method to spatially express the US and the EM systems within the same coordinate system. This calibration consisted in computing the transformation matrix between the US coordinate system and the EM coordinate system. Once this calibration matrix was computed, US and EM data can be expressed into the same coordinate system as long as the sensor remains fixed on the US probe. Such a technique allows any movements of the head or the transducer to be compensated for, and data can be hence collected without a dedicated support system [6].

4.3. Temporal synchronization

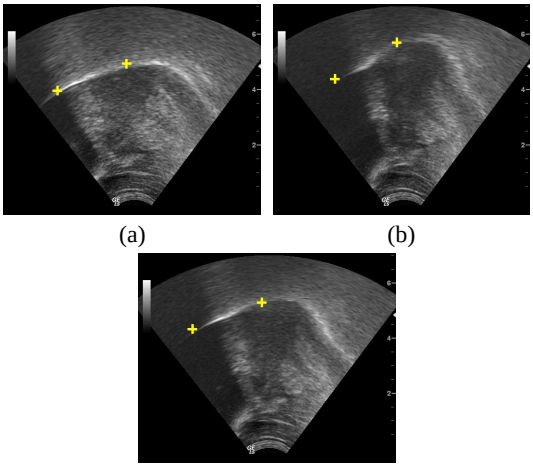
During each acquisition, one beep is emitted at the beginning and at the end by the control PC. These two beeps provide an es-

timation of the sampling rate of each modality (audio and EM) as explained in Section 3.

4.4. Results

The corpus was tested with success on the speaker. The two sensors on the tongue moved accordingly to the tongue shape onto US images, showing that both spatial calibration and temporal synchronization were satisfactory. Moreover, the sound was coherent with the movement. This was especially visible on motions where the movement of the tongue shape was moving fast, such as */f/* for example (Fig. 6).

In order to visually check the temporal synchronization, complete sequences mixing the three modalities (US, EM and audio) are available on our website at http://magrit.loria.fr/Confs/Interspeech07. One video with a lower definition is attached to this paper.



(a) (b)

(c)

Figure 6: *EM points (crosses) onto US tongue images. Apex is on the left side of the image: (a) first /i/ from /ifi/. (b) /f/ from /ifi/. (c) second /i/ from /ifi/.*

5. Conclusions

This paper presents a complete setup for coupling EM sensors, US images and audio data to get automatically synchronized video sequences of the tongue during speech. Each modality keeps its sampling rate without resampling the data. This setup has been successfully used to record a corpus that will be used to elaborate more precise acoustic models of speech production and to evaluate audiovisual-to-articulatory inversion methods.

The other key point of our system is the calibration of geometrical modalities (ultrasound, electromagnetic sensors and stereovision). In addition to enabling all these data to be exploited together this avoids the contention of the subject, and the ultrasound probe to be fixed. Speech articulation is consequently closer to natural speech even if there are still two wires for the electromagnetic sensors.

For clarity sake, and because most of the synchronization problems were covered by the modalities mentioned above, we have only described the coupling between electromagnetic sensors, ultrasound imaging and audio. Actually, our synchronization framework can be applied to other modalities, such as stereovision cameras to get the speaker's face shape. The principle remains identical, with a control PC that controls LEDs (events visible within the camera fields) and audio beeps. Video images can thus be synchronized with the sound, and therefore with other data acquired. The first corpus recorded by using our system comprises the acoustical speech signal, ultrasound images, 3D positions of the electromagnetic sensors, and stereo images enabling the tracking of face deformations (see [9] for further details about the stereovision system used).

Future improvements will concern the US synchronization: the present delay accuracy corresponds to ± 2 US images. This gives an accurate visual synchronization between US and other data. However the determination of the exact image offset could probably improve the visual impression for fast speech sequences. We plan to compute this offset by comparing the movement given by the optical flow computed on ultrasound images with that of the electromagnetic sensors.

6. Acknowledgements

This work is part of the ASPI project funded by the IST Programme of the Commission of the European Communities as project number IST- 2005-021324. We would like to acknowledge people from NDI, Germany, for their help and support.

7. References

[1] Guenther, F.H. and Perkell, J.S. A neural model of speech production and its application to studies of the role of auditory feedback in speech, *Speech Motor Control in Normal and Disordered Speech*, pp 29–49, 2004

[2] Rose, R.C., Schroeter, J. and Sondhi, M.M. An investigation of the potential role of speech production models in automatic speech recognition *Proc. of Int. Conf. on Spoken Language Processing (ICSLP'94)*, 2, pp 575–578, 1994

[3] Fontecave, J. and Berthommier, F. Quasi-automatic extraction method of tongue movement from a large existing speech cineradiographic database *Interspeech 2005*, 2005

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

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

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

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

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

[9] Wrobel-Dautcourt B., Berger M.O., Potard B., Laprie Y. and Ouni S.: A low cost stereovision based system for acquisition of visible articulatory data, *in Proc. of Int. C. on Auditory-Visual Speech Processing (AVSP'05)*, pp 145–150, 2005.