Acquisition of Ultrasound, Video and Acoustic Speech Data for a Silent-Speech Interface Application

T. Hueber[1,3], G. Chollet[3], B. Denby[2,1], M. Stone[4]

[1]Laboratoire d'Electronique, ESPCI ParisTech, Paris, France
[2]Université Pierre et Marie Curie – Paris VI, Paris, France
[3]Laboratoire Traitement et Communication de l'Information, Telecom ParisTech, Paris, France
[4]Vocal Tract Visualization Lab, University of Maryland Dental School, Baltimore, USA

hueber@ieee.org, gerard.chollet@tsi.enst.fr, denby@ieee.org, mstone@umaryland.edu

Abstract

This article addresses synchronous acquisition of high-speed multimodal speech data, composed of ultrasound and optical images of the vocal tract together with the acoustic speech signal, for a silent speech interface. Built around a laptop-based portable ultrasound machine (Terason T3000) and an industrial camera, an acquisition setup is described together with its acquisition software called Ultraspeech. The system is currently able to record ultrasound images at 70 fps and optical images at 60 fps, synchronously with the acoustic signal. An interactive inter-session re-calibration mechanism which allows recording of large audiovisual speech databases in multiple acquisition sessions is also described.

1 Introduction

Measuring the activity of the vocal tract during speech is critical in a variety of fields such as phonology, linguistics, speech pathology, anatomy and multimodal speech processing. In recent years and despite the success of MRI, the use of ultrasound for vocal tract imaging and analysis remains popular mainly because of its non-invasive property, its good time resolution, its clinical safety, and its ability to image the tongue in non-supine subjects.

In the "silent speech interface" developed in the *Ouisper* project, non-acoustic features, derived from ultrasound images of the tongue and optical images of the lips, are used to drive a speech synthesizer. A laryngectomyzed patient could use this kind of system as an alternative to oesophageal speech, tracheo-oesophageal speech or the electrolarynx. A silent-speech interface could also be useful in situations where silence must be maintained, or for voice communication in noisy environments. Described in [1] and [2], the *Ouisper* segmental speech coder is built around a large audiovisual corpus (more than one hour) which associates articulatory features extracted from a 30 Hz source video with acoustic observations. In the proposed approach, a sequence of phones is "recognized" from visible motion of the tongue and lips. In the visuo-phonetic decoding stage, phonetic prediction is hampered by a large number of deletion errors. Most often, deleted phones are very short phones for which correct visualization is difficult with a 30 Hz acquisition system. Thus, a faster acquisition system is needed.

Several systems able to acquire a sequence of ultrasound images of the tongue together with the uttered speech signal have been described in the literature. However, the coupling of an ultrasound imaging system with another imaging device, such as a high-speed camera, without decreasing the acquisition framerate, remains a difficult problem. In this article, a new acquisition system is introduced, which in addition to the acoustic signal, is able to synchronously record both ultrasound and optical streams at more than 60 fps on a single and "easy-to-transport" laptop-based machine.

Section 2 of the article presents briefly the state-of-the-art in ultrasound speech data acquisition via a

non-exhaustive list of existing solutions. Both hardware and software components of the proposed acquisition system, which is based on the *Terason T3000* portable ultrasound system coupled with an industrial camera and driven by the dedicated *Ultraspeech* application, are described in section 3. Finally, the capacity of the system to record synchronously multiple high-speed data streams is evaluated experimentally in section 4.

2 State of the Art in Multimodal Speech Data Acquisition System

Much of the progress that has been achieved in multimodal speech data acquisition systems concerns the positioning of the head and the transducer: These may be stabilized as in HATS [3] or related systems [4]; free to move using a helmet arrangement [5]; or tracked, using infrared emitting diodes (HOCUS [6]), visible markers (PALATRON [7]) or electromagnetic sensors [8].

The other crucial issue in such acquisition systems is the synchronisation of the ultrasound image stream with the uttered acoustic speech signal. In most systems ([3], [4], [6], [7]), this task is performed using an analog video mixer which downsamples the ultrasound data stream to 30 Hz. In the system proposed by Aron [8], ultrasound, electromagnetic and audio data are recorded synchronously with each modality keeping its own framerate. The synchronization of ultrasound data with audio is achieved in that system by time-aligning the ultrasound machine cineloop (a video buffer of the last 15 seconds recorded) with a timecode on an external PC.

3 Description of the Acquisition System

In the context of a silent speech interface based on tongue and lip imaging, the desired acquisition system should be able to record synchronously ultrasound data and video data at their respective maximum framerate together with the acoustic speech signal. In order to have a compact, transportable, and easy-to-use system, a PC-based hardware architecture coupled with a single control program has been adopted.

3.1 Hardware component of the system

As shown in figure 1, the hardware component of the system is based on:

- the *Terason T3000* ultrasound system which is based on a laptop running Microsoft Windows XP and provides 640x480 pixels resolution images
- a 140° microconvex transducer with 128 elements (8MC4)
- an industrial USB color camera able to provide 60 fps with a 640x480 pixels resolution (USB 2.0, WDM compliant)
- an external microphone connected to the built-in soundcard of the *T3000*

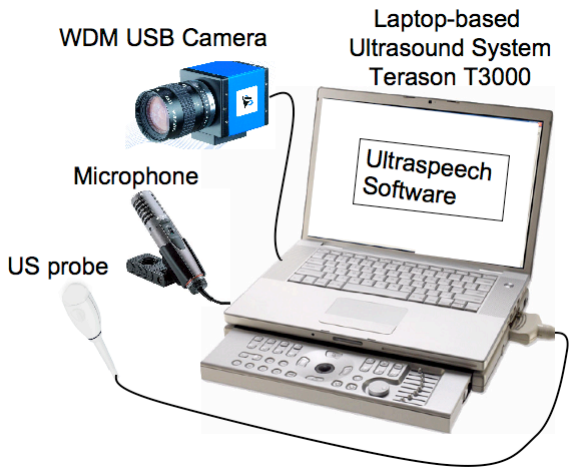

Figure 1: *Hardware component of the acquisition system*

In the described system, data streams are recorded, processed and stored digitally on a single PC using our stand-alone software *Ultraspeech*.

3.2 The *Ultraspeech* software

The open shared-memory client-server architecture of the *Terason T3000* system allows the development of stand-alone client applications with real-time access to the live stream of ultrasound images. The *Ultraspeech* MFC application (*Microsoft Foundation Classes*) is optimized for the *Terason T3000*, supports WDM compliant cameras (*Windows Driver Model*), and DirectX compatible soundcards. As shown in figure 2, *Ultraspeech* allows the real-time visualization of image streams and the automation of

the imaging devices. Internally, *Ultraspeech* uses multiple FIFO buffers to access image data.

Figure 2: *Ultraspeech software (main window) with ultrasound and camera visualization area (a and b), text stimuli display area (c), console output (d) and start/stop recording button (e)*

The main feature of the *Ultraspeech* software is the synchronous recording of both image streams with the audio signal. Data recording is triggered by simply clicking on a start/stop button. Thanks to multithreading programming techniques, all streams are processed in parallel. Streams share the same multimedia timer so that each frame and each audio buffer can be tagged with the timer value during the recording. Any initial asynchrony between streams is captured during the acquisition, and synchrony is restored automatically in a post-processing stage. The entire recording procedure is fully automatic and no *a posteriori* human check is needed. After each acquisition, data are directly available as series of bitmaps for both image streams and WAV files for the audio stream, in the specified directory (local or remote). Furthermore, *Ultraspeech* provides convenient tools for large database recording, such as an automatic file naming system and the automatic display of the text stimuli for each item to record (*i.e* the word or sentence to pronounce).

3.3 Inter-session re-calibration procedure

Techniques involved in the silent vocoder described in [1] require the recording of a large amount of multimodal speech data. In our earlier work, data was recorded in a single long session during which the subject remained fixed in the HATS system. Data acquisition in multiple sessions (spaced in time) requires an inter-session re-calibration mechanism to position the speaker's head at a reference position (the probe remains fixed). The procedure shown in figure 3 is based on real-time averaging of a live image with a target reference image. During this interactive re-calibration procedure, the subject adjusts the position of his/her head in order to fit to the target reference position. A similar procedure is used for ultrasound, where the live tongue image is super-imposed on a target reference. When coupled with a head stabilization system, this procedure is convenient, rapid and effective.

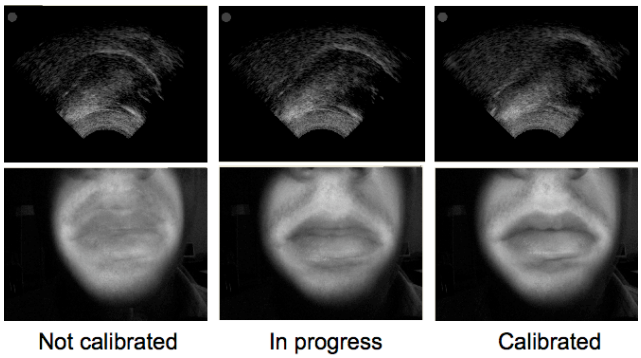

Figure 3: *Interactive speaker inter-session re-calibration mechanism at different stages of the procedure*

4 Experimental results

In order to check the synchronization of the different streams, the following experiment, illustrated in figure 4, is done. A hammer is used to tap a bottle of ultrasound gel (fig 4, a and b), ejecting a droplet onto the probe (c). The droplet shows up immediately on the probe (e) and should be synchronized with video of the hammer (d) hitting the bottle and the sound of this contact (f). Stream synchrony can be observed in figure 4 where a 71 fps ultrasound stream is displayed with a 60 fps video stream and the audio signal on the same time scale.

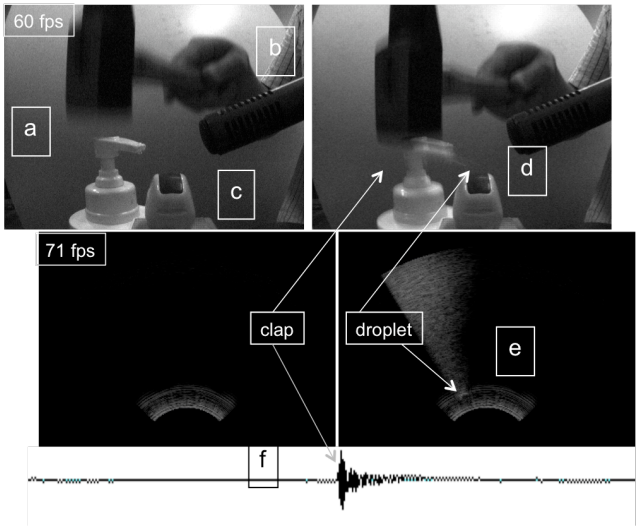

Figure 4: Interactive *speaker inter-session re-calibration mechanism (ultrasound is clearly synchronized with video and audio)*

This synchronization check procedure has been shown to be highly reproducible and has been tested on a variety of different ultrasound probes (*Terason 5MC2, 8MC3, 8MC4*). The ultrasound and video streams are found always to be synchronized. A residual delay occasionally observed between visual (ultrasound and video) and audio is always less than the inter-frame gap (*i.e* 15 ms at 60 fps). To summarize, the system is currently able to record synchronously:

- the ultrasound stream at **71 fps** (7cm depth, 320x240 pixels resolution, bitmap format)
- the video stream at **60 fps** (640x480 pixels resolution, bitmap format)
- the audio signal (44100 Hz, 16 bits, mono, PCM)

5 Conclusion and Perspectives

The conception of a silent speech interface based on tongue and lip imaging requires a high-speed acquisition system. The flexible PC-based architecture of the *Terason T3000* ultrasound system has allowed the development of *Ultraspeech*. This software interfaces ultrasound with a high-speed USB camera and an audio device. The system is able to record synchronously these different data streams which retain their respective framerates. For the recording of large databases in multiple acquisition sessions, an inter-session re-calibration procedure has also been introduced. The system has been used for the recording of small databases (100 words) and is now ready to be validated on a large dataset recording task.

6 Acknowledgements

This work is supported by the French Department of Defense (DGA) and the French National Research Agency, under contract number ANR-06-BLAN-0166.

7 References

[1] T. Hueber, G. Chollet, B. Denby, G. Dreyfus, M. Stone, “Towards a Segmental Vocoder Driven by Ultrasound and Optical Images of the Tongue and Lips”, Interspeech, to appear, Brisbane, Australia, 2008.

[2] T. Hueber, G. Chollet, B. Denby, M. Stone, L. Zouari, “Ouisper: Corpus Based Synthesis Driven by Articulatory Data”, International Congress of Phonetic Sciences, Saarbrücken, pp. 2193-2196, Germany, 2007.

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

[4] L. Davidson, “Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance”, Journal of the Acoustical Society of America 120:1, 407-415, 2005.

[5] A. Wrench, J. Scobbie, M. Linden, “Evaluation of a helmet to hold an ultrasound probe”, Ultrafest IV, NY, USA, 2007.

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

[7] J. Mielke, A. Baker, D. Archangeli, S. Racy, “Palatron: A Technique for Aligning Ultrasound Images of the Tongue and Palate”, in D. Siddiqi, and B. V. Tucker, Eds., Coyote Papers. vol. 14. 97-108, 2005.

[8] M. Aron, N. Ferveur, E. Kerrien, M.O. Berger, Y. Laprie, “Acquisition and synchronization of multimodal articulatory data”, Interspeech, Antwerp, Belgium, 2007.