Ultraspeech, a portable system for acquisition of high-speed ultrasound, video and acoustic speech data

T. Hueber¹,³, G. Chollet³, B. Denby²,¹

¹Laboratoire d'Électronique, ESPCI ParisTech, Paris, France
²Université Pierre et Marie Curie – Paris VI, Paris, France
³Laboratoire Traitement et Communication de l’Information, Telecom ParisTech, Paris, France

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

The article introduces the Ultraspeech acquisition system, which allows to synchronously acquire high-speed ultrasound and video images of the vocal tract together with the acoustic speech signal. The system has been primarily designed for a “silent speech interface”, in which non-acoustic features, derived from ultrasound and video images of the tongue and lips, are used to drive a speech synthesizer. Ultraspeech is built around and optimized for two imaging devices: the lightweight, portable ultrasound machine Terason T3000 (laptop or Firewire OEM version) and a 60 fps USB WDM-compliant industrial camera (from Imaging Source). The originality of our approach resides in the fact that the ultrasound, video and sound devices are treated as external multimedia sources connected via standard communication bus to a standard (Microsoft) Windows-based laptop. This architecture makes the system compact, transportable (when battery-powered), and easy-to-use. The imaging devices as well as the sound system are controlled by a stand-alone dedicated Ultraspeech software application package.

The latter is a “user friendly” graphical (Windows-like) application that allows synchronous recording of the two image streams and the audio signal, at their respective maximum frame rates. Using multithreading programming techniques, the visual and audio streams can be processed in parallel. These streams share a common timer so that each frame and audio buffer can be tagged with the timer value while recording. Any initial asynchrony between streams is captured during the acquisition, allowing synchrony to be restored automatically in a post-processing stage. The entire recording procedure is fully automatic (in push-button fashion) and no a posteriori human checking is required. After each acquisition, data are directly available as series of bitmaps for the image streams, and WAV files for the audio, in a previously specified directory. Ultraspeech also provides convenient tools for recording large databases, such as an automatic file naming system, the automatic display of the text stimuli, and an interactive inter-session re-calibration mechanism that allows acquisition of large audiovisual speech databases over multiple acquisition sessions.

A simple manual procedure similar to a “cinema clap” is used to check the synchrony of the recorded streams. Experimental results show that the residual delay between visual (ultrasound and video) and audio streams is always less than the inter-frame gap (i.e., 15 ms at 60 fps). With a focal distance at 7 cm, appropriate for tongue visualization, the system is currently able to record simultaneously and synchronously: the ultrasound stream at 70 fps (image resolution of 320x240 pixels); the video stream at 60 fps (640x480 pixels); and the acoustic signal at 16 kHz.

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