Ultraspeech-player: Intuitive visualization of ultrasound articulatory data for speech therapy and pronunciation training

Thomas Hueber

GIPSA-lab, UMR 5216/CNRS/INP/UJF/U.Stendhal, Grenoble, France
thomas.hueber@gipsa-lab.grenoble-inp.fr

Abstract

This paper introduces *Ultraspeech-player*, a software dedicated to the visualization of ultrasound and video sequences of the tongue and lips. This software is designed for speech therapy and pronunciation training applications and aims at increasing the articulatory awareness of the learner. *Ultraspeech-player* includes an audiovisual time-stretching algorithm allowing the user to slow-down in real-time both the articulatory gesture and its corresponding acoustic realization. This rendering technique aims at improving the way a naïve speaker perceives and understands an articulatory gesture.

Index Terms: ultrasound, speech therapy, CAPT

1. Introduction

Several studies tend to show that the visualization of the articulatory movements facilitates pronunciation training [1]. In recent years, the use of ultrasound imaging remains popular to capture tongue movements during speech. Ultrasound imaging is a non-invasive technique which has a good time resolution, is clinical safe, and can be used to image the tongue in non-supine subjects. In [2], we introduced *Ultraspeech* (www.ultraspeech.com) a standalone software allowing the synchronous and simultaneous acquisition of high-speed ultrasound, video, and audio signals, at their respective maximum temporal resolution. *Ultraspeech* has been primarily designed for a silent speech interface, in which ultrasound and video images of the tongue and lips were used to drive a speech synthesizer [3]. In this paper, we introduce *Ultraspeech-player*, a standalone software dedicated to the visualization of ultrasound speech data recorded using *Ultraspeech*. *Ultraspeech-player* is designed for pronunciation training in the context of speech therapy and second language learning. The software aims at displaying natural tongue movements acquired on a reference speaker, for different kind of sequences (isolated vowels, VCV, swallowing, etc.). It also includes an audiovisual time-stretching module allowing the user to slow-down both the articulatory gesture and its corresponding acoustic realization (i.e. the speed of the audio signal is modified while the original pitch it preserved). This rendering technique aims at improving the way a naïve speaker perceive and understand a tongue gesture.

2. Software architecture and features

2.1. Architecture

Ultraspeech-player is composed of:

- a standalone software embedding the main graphical user interface (developed using the Max/MSP programming environment) and the audiovisual time-stretching module (written in C and wrapped as a *Max/MSP external*).
- a set a databases, recorded on a *reference speaker* using the *Ultraspeech* system.

Each database contains:

- a set of high-speed ultrasound and video sequences of tongue and lips movements.
- The corresponding set of audio files encoded using an Harmonic Plus Noise (HNM) modeling technique [4].
- a dedicated sub graphical user interface (sub-GUI) adapted to the type of data (phonemes, syllables, words, etc). This sub-GUI is incrusted automatically in the main GUI when the database is loaded.

A screenshot of *Ultraspeech-player* is presented at Figure 1.

2.2. Default databases

In its current (beta) version (v0.1), *Ultraspeech-player* includes 3 databases recorded on a female French native speaker. These databases contain respectively all the French vowels, all the French consonants (pronounced in an isolated manner) and all the French VCV (Vowel-Consonant-Vowel) sequences. Ultrasound images of the tongue were recorded in the mid-sagittal plane using the Terason T3000 ultrasound system with a 140° angle microconvex ultrasound probe (7cm imaging depth, 71 fps). Video sequences of the lips were recorded using a 60 fps industrial CMOS camera. Ultrasound probe and video camera were attached to the speaker's head using an adapted version of the *Articulate Instrument* headset (www.articulateinstruments.com). The audio signal was recorded in an anechoic room and sampled at 44100 Hz (32 bits). All the streams were recorded simultaneously and synchronously using the *Ultraspeech* software. The data acquisition setup is presented at Figure 2.

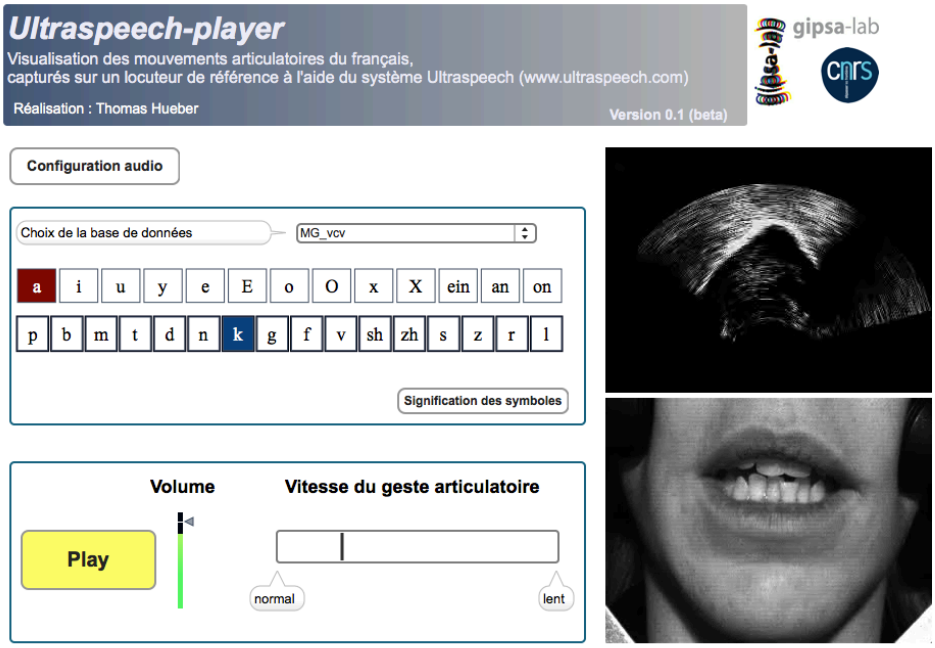

Figure 1: *Ultraspeech-player 0.1 (beta) screenshot.*

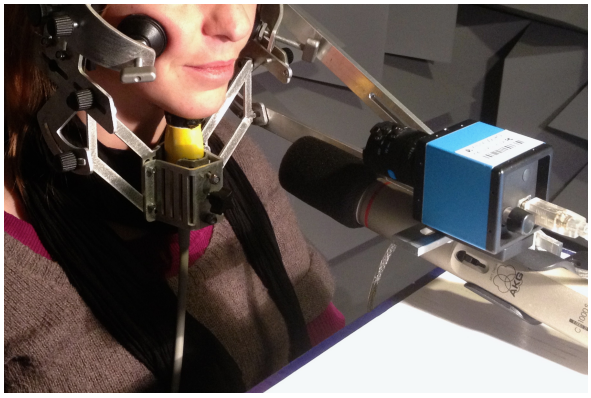

Figure 2: *Data acquisition setup.*

2.3. Audiovisual time-stretching

Ultraspeech-player embeds a time-stretching module allowing the user to control the speed of the displayed articulatory gesture, in real-time. Audio processing is achieved using a real-time implementation of the HNM vocoder. This technique allows adjusting the length of an audio signal without modifying its pitch. Stretching factor is controlled via the software GUI using a basic slider. The playing speed of the visual data (ultrasound and video) is adjusted in real-time in order to match the desired stretching factor.

3. Conclusions

Ultraspeech-player allows the visualization of tongue and lips movements during speech, captured on a reference speaker using ultrasound and video imaging. In future work, we will focus on improving the display of the tongue movements, by adding to the ultrasound image: the palatal trace, the teeth, and a contour of the phararyngeal cavity. This data will be extracted from MRI anatomical scans from the same speaker. *Ultraspeech-player* v0.1 is free to download at www.ultraspeech.com/player and is currently experimented by a group of 5 speech therapists in Grenoble (France). Latest release of *Ultraspeech-player* will be demonstrated during *Show&Tell* event of Interspeech 2013.

4. Acknowledgments

The author would like to thank Audrey Acher, Marine Verdurand and Amandine Gros for their useful suggestions, and the speaker MG for her help in recording the articulatory databases.

5. References

[1] Badin, P., Ben Youssef, A., Bailly, G., Elisei, F., and Hueber, T., "Visual articulatory feedback for phonetic correction in second language learning", in Proc. of SLATE workshop, P1-10, 2010.

[2] Hueber, T., Benaroya, E.L., Chollet, G., Denby, B., Dreyfus, G., Stone, M., (2009) "Development of a Silent Speech Interface Driven by Ultrasound and Optical Images of the Tongue and Lips", Speech Communication, 52(4), pp. 288-300..

[3] Hueber, T., Chollet, G., Denby, B., and Stone, M. (2008). "Acquisition of ultrasound, video and acoustic speech data for a silent-speech interface application," Proceedings of International Seminar on Speech Production (Strasbourg, France), pp. 365-369.

[4] Stylianou, Y., Dutoit, T., Schroeter, J., "Diphone Concatenation using a Harmonic plus Noise Model of Speech," Eurospeech, pp. 613-616, Rhodes, Greece, 1997.