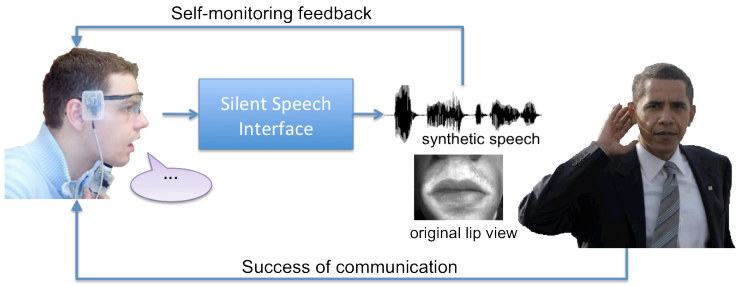
Available PhD position with scholarship - Silent speech interface
GIPSA-lab, Grenoble, France

Incremental speech synthesis for a real-time silent speech interface

Context: The design of a *silent speech interface*, i.e. a device allowing speech communication without the necessity of vocalizing the sound, has recently received considerable attention from the speech research community [1]. In the envisioned system, the speaker articulates normally but does not produce any audible sound. Application areas are in the medical field, as an aid for larynx-cancer patients, and in the telecommunication sector, in the form of a "silent telephone", which could be used for confidential communication, or in very noisy environments. In [2], we have shown that ultrasound and video imaging can be efficiently combined to capture the articulatory movements during silent speech production; the ultrasound transducer and the video camera are placed respectively beneath the chin and in front of the lips. At present, our work focused mainly on the estimation of the target spectrum from the visual articulatory data (using artificial neural network, Gaussian mixture regression and hidden Markov modeling). The other challenging issue concerns the **estimation of acceptable prosodic patterns** (i.e. the *intonation* of the synthetic voice) from silent articulatory data only. To address this ill-posed problem, one solution consists of splitting the mapping process into two consecutive steps: (1) a visual speech recognition step which estimates the most likely sequence of word given the articulatory observations, and (2) a text-to-speech (TTS) synthesis step which generates the audio signal from the decoded word sequence. In that case, the target prosodic pattern is derived from the linguistic structure of the decoded sentence. The major drawback of this mapping method is that it cannot run in real-time. In fact, if the visual speech recognition step can be done *online* (i.e. words are decoded a short amount of time after they have been pronounced), standard TTS systems need to know the entire sentence to estimate the target prosody. This introduces a large delay between the (silent) articulation and the generation of the synthetic audio signal. This delay prevents the communication partners from having a fluent conversation. The main goal of this PhD project is to design a **real-time silent speech interface**, in which the delay between the articulatory gesture and the corresponding acoustic event has to be constant and as short as possible.

Goals: The goal of this PhD project is twofold:

(1) Reducing the delay between the recognition and the synthesis steps, by designing a new generation of TTS system, called "**incremental TTS system**" [3]. This system should be able to synthesize the decoded words, with acceptable prosody, as soon as they are provided by the visual speech recognition system.



(2) Designing experimental paradigms in order to evaluate the system in realistic communication situations (face-to-face, remote/telephone-like interaction, human-machine interaction). The goal is to study how a silent speaker benefits from the acoustic feedback provided by the incremental TTS and how he/she adapts his/her own articulation to maximize the efficiency of the communication.

Supervision: Dr. Thomas Hueber, Dr. Gérard Bailly (CNRS/GIPSA-lab)

Duration / Salary: 36 months (October 2013- October 2016) / ~ €1400/month minimum (net salary).

Research fields: multimodal signal processing, machine learning, interactive systems, experimental design

Background: Master's or engineer's degree in computer science, signal processing or applied mathematics.

Skills: Good skills in mathematics (machine learning) and programming (Matlab, C, Max/MSP). Knowledge in speech processing or computational linguistics would be appreciated.

To apply: send your CV, transcript of records of your Master grade and a cover letter to thomas.hueber@gipsa-lab.grenoble-inp.fr

References:

[1] B. Denby, T. Schultz, K. Honda, T. Hueber, et al., "Silent Speech Interfaces," Speech Communication, vol. 52, no. 4, pp. 270-287, 2010.

[2] T. Hueber, E. L. Benaroya, G. Chollet, et al., "Development of a Silent Speech Interface Driven by Ultrasound and Optical Images of the Tongue and Lips", Speech Communication, vol. 52, no. 4, pp. 288-300, 2010.

[3] Buschmeier H, Baumann T, Dosch B, Schlangen D, Kopp S. "Combining Incremental Language Generation and Incremental Speech Synthesis for Adaptive Information Presentation", in proc of the 13th Sigdial meeting, pp, 295-303, 2012.