Proposal for a phd preparation position at GIPSA-lab (team CRISSP)

Title - Modelling for Augmented Speech Lip-Reading: audio-visual conversion and visual recognition

Thesis advisor: Dr. Denis Beautemps

Co-supervisor: Professor Gang Feng

Contact: Dr. Denis Beautemps (Denis.Beautemps@gipsa-lab.grenoble-inp.fr)

Salary: According to French standards (about 1400 euros net/month)

Skills: We are looking for an open-minded student, coming from an engineering background (computer science, mathematics, signal and data processing)

Closing date for application: Send a CV with a transcript of higher education records (at Master and postgraduate level) and the rank before 10 June 2015

Description of the project:
The aim of this project is the automatic conversion of Speech (i) in the visual modality from the audio modality and inversely (ii) in the audio modality from the visual modality. The visual modality eventually augmented by Cued Speech (Fig. 1) is used for the communication with deaf people (the so-called “lip-reading”). One needs conversion between the audio and visual spaces for communication between hearing people who use principally the speech audio modality and the deaf people orally educated who visual is the main modality for speech perception. The methods used for the conversion can be ordered on an axis from the mapping methods between the parameters of the two spaces i.e. at a signal level of interface to the full phonetic recognition i.e. at a symbol level through methods combining mapping and classification (see Fig. 2 for an illustration).
The project will need first to collect large audio-visual corpus made of phonemes, syllables, isolated words and sentences in French language uttered and cued in Cued Speech by multiple subjects. The utterances will be recorded in video. The GIPSA-lab has an anechoic room equipped with cameras. This step will be completed by the extraction of parameters from the recordings in order to build a database used for the modelling. The parameters carrying the language traces will be considered. Typically, the audio space is defined by spectral components derived from the audio speech. The visual space is described by parameters derived from the lip region of interest and related to the lip articulation (shape or appearance features) combined with parameters of the hand region of interest eventually completed by additional parameters derived from other regions of interest involved in the interaction.
A second important step of the work will be the automatic selection of the relevant images and inside each selected images the segmentation of the different regions of interest (the face, lips, hand,…) and their association to take into account their possible de-synchronization (the hand in Cued Speech is often in advance on the lips). Learning methods, recognition tools used in Image processing and more generally in computer vision could be used to locate the ROI and to follow the objects. These latter could be combined with tools used for processing collision objects as it is the case between hand and face for Cued Speech. The pertinence and the choice of the parameters will be discussed in function of their dimensionality and their robustness to the recording constraints (brightness, face-to-face situations or not …).
The results of this first part will allow to open two other ones in the domain of modelling. The first one will focus on the conversion of the audio towards the image made of the face, lips and hand. The results obtained at the GIPSA-lab for the mapping of hand placement and lip parameters in the case of vowels (Ming, Beautemps, Feng, 2013) will be a basis for the extension to the other speech units. The hand movement will require particular attention for the modelling. Indeed, the hand movement is made of a transition between hand placement locations and plateaus at the target hand placement location. The formation of the hand configuration is super-imposed to the transition phase. Except its duration, the trajectory of the hand during the transition has no evident relation with speech production. Such a system requires to be used from the point of view of modelling to resort to particular methods using the principles of learning machine and/or fuzzy logic.

The last part will focus on the conversion of lip parameters augmented by Cued Speech ones in the context of visual towards audio conversion. The mapping towards the spectral parameters and speech recognition will be concerned. The approach consists to extend our previous work on vowels (Aboutabit thesis, 2007; Heracleous, Beautemps, Aboutabit, 2010) to all the speech units. More precisely the fusion methods of parameters will be explored. The approach in direct fusion, or in fusion of decisions, or hybrid as ordered by Potamianos et al (2012) will be evaluated.

This research program is ambitious but will be further refined in function of the profile candidate.
This program takes place in the context of automatic lip-reading, domain that regains in popularity, not only for scientific purpose but also for great interest in socio-economical context and in particular in deaf rehabilitation and security/surveillance due to the increase in calculation possibilities of the computers. The last results in lip reading alone recognition systems attain the accuracy of 76% (recent works of Prof. Ahmad Hassanat of the Mu’tah, University of Jordan) but correspond to similar performances reached since 2010 at GIPSA-lab (76.1% for the best ones, see Heracleous, Beautemps, Aboutabit, 2010). It has to be recall that this result is way above the performances obtained naturally with the best human lip-readers. Scores better than 76% can be obtained with the addition of supplementary information such as language models or Cued Speech codes for instance. We could thus obtain an accuracy of 94.9% with Cued Speech on a vocabulary of isolated words (Heracleous, Beautemps, Aboutabit, 2010), performances similar to the results in speech recognition with the audio signal.

