Speech Unit(e)s
The multisensory-motor unity of speech

Understanding how speech unites the sensory and motor streams,
And how speech units emerge from perceptuo-motor interactions

ERC Advanced Grant, Jean-Luc Schwartz, GIPSA-Lab, Grenoble, France

Proposal for a post-doc position: "The informational structure of audio-visuo-motor speech"

Context

The ***Speech Unit(e)s*** project is focused on the speech unification process associating the auditory, visual and motor streams in the human brain, in an interdisciplinary approach combining cognitive psychology, neurosciences, phonetics (both descriptive and developmental) and computational models. The framework is provided by the "Perception-for-Action-Control Theory (PACT)" developed by the PI (Schwartz et al., 2012).

PACT is a perceptuo-motor theory of speech communication connecting, in a principled way, perceptual shaping and motor procedural knowledge in speech multisensory processing. The communication unit in PACT is neither a sound nor a gesture but a perceptually shaped gesture, that is a perceptuo-motor unit. It is characterized by both articulatory coherence – provided by its gestural nature – and perceptual value – necessary for being functional. PACT considers two roles for the perceptuo-motor link in speech perception: online unification of the sensory and motor streams through audio-visuo-motor binding, and offline joint emergence of the perceptual and motor repertoires in speech development.

Objectives of the post-doc position

In a multisensory-motor context such as the present one, a major requirement concerns a better knowledge of the structure of the information. Speech scientists have acquired a very good knowledge of the structure of speech acoustics, capitalizing on large audio corpora and up-to-date statistical techniques (mostly based on sophisticated implementations of Hidden Markov Models, e.g. Schutz & Waibel, 2001, Yu & Deng, 2011). Data on speech rhythms, in relation with the syllabic structure, have been analyzed clearly in a number of works (Greenberg, 1999; Grant & Greenberg, 2003).

In spite of strong efforts in the field of audiovisual speech automatic recognition (Potamianos et al., 2003), characterization of the structure of audiovisual information is scarce. While an increasing number of papers on audiovisual speech perception presenting cognitive and neurophysiological data quote the "advance of the visual stream on the audio stream", few papers provide quantitative evidence (see Chandrasekaran et al., 2009) and when they do, these are sometimes mistaken or oversimplified. Actually, the temporal relationship between visual and auditory information is far from constant from one situation to another (Troille et al., 2010). Concerning the perceptuo-motor link, the situation is even worse. Few systematic quantitative studies are available because of the difficulty to acquire articulatory data, and in these studies the focus is generally set on the so-called "inversion problem" (e.g. Ananthakrishnan & Engwall, 2011, Hueber et al., 2012) rather than a systematic characterization of the structure of the perceptuo-motor relationship. Finally, there is a strong lack of systematic investigations of the relationship between orofacial and bracchio-manual gestures in face-to-face communication.

We shall gather a large corpus of audio-visuo-articulatory-gestural speech. Articulatory data will be acquired through ultrasound, electromagnetic articulography (EMA) and video imaging. Labial configurations and estimates of jaw movements will be automatically extracted and processed thanks to a large range of video facial processing. We also additionally plan to record information about accompanying coverbal gestures by the hand and arm, thanks to an optotrack system enabling to track bracchio-manual gestures. A complete equipment for audio-video-

ultrasound-EMA-optotrack acquisition and automatic processing, named ultraspeech-tools, is available in Grenoble (more information on www.ultraspeech.com). The corpus will consist in isolated syllables, chained syllables, simple sentences and read material. Elicitation paradigms will associate reading tasks (with material presented on a computer screen) and dialogic situations between the subject and an experimenter to evoke coverbal gestures as ecologically as possible.

The corpus will be analyzed, by extensive use and possibly development of original techniques, based on data–driven statistical models and machine learning algorithms, in the search for three major types of characteristics:

1) ***Quantification of auditory, visual and motor rhythms***, from kinematic data – e.g. acoustic envelope, lip height and/or width, variations in time of the principal components of the tongue deformations, analysis of the arm/hand/finger dynamics;
2) ***Quantification of delays*** between sound, lips, tongue and hand in various kinds of configurations associated with coarticulation processes (e.g. vowel to vowel anticipatory and perseverative phenomena, consonant-vowel coproduction, vocal tract preparatory movement in silence) and oral/gestural coordination;
3) ***Quantification of the amount of predictability*** between lips, tongue, hand and sound, through various techniques allowing the quantitative estimate of joint information (e.g. mutual information, entropy, co-inertia) and perform statistical inference between modalities (e.g. Graphical models such as dynamic Bayesian Framework, multi-stream HMM, etc.)

The work will be performed within **a multidisciplinary group in GIPSA-Lab Grenoble, associating specialists in speech and gesture communication, cognitive processes, signal processing and machine learning** (partners of the project: Jean-Luc Schwartz, Marion Dohen from the "Speech, Brain, Multimodality Development" team; Thomas Hueber, Laurent Girin from the "Talking Machines and Face to Face Interaction" team; Pierre-Olivier Amblard and Olivier Michel from the "Information in Complex Systems" team).

Practical information

The post-doc position is open for a two-year period, with a possible third-year prolongation.

Candidates should have a background in speech and signal processing, face-to-face communication, and machine learning, or at least two of these three domains.

Candidates should send a short email to Jean-Luc Schwartz (Jean-Luc.Schwartz@gipsa-lab.grenoble-inp.fr) to declare their intention to submit a full proposal.

Then they must send a full application file which will include an extended CV and a list of publications, together with a letter explaining why they are interested in the project, what their specific interests could be, possibly suggesting other experiments related to the general question of the informational structure of audio-visuo-motor speech, and also how this position would fit into their future plans for the development of their own career. They should also provide two names (with email addresses) for recommendations about their applications. Preselected candidates will be interviewed.