2018-00654 - PhD Position /Studying gestures and speech for an effective robot-human interaction [S]

Level of qualifications required: Graduate degree or equivalent
Other valued qualifications: Master of computer science
Function: PhD Position

Context
One of the main objectives of social robotics research is to design and develop robots that can engage in social environments in a way that is appealing and familiar to humans. However, interaction is often difficult because users do not understand the robot’s internal states, intentions, actions, and expectations. Thus, to facilitate successful interaction, social robots should provide communicative functionality that is both natural and intuitive. Given the design of humanoid robots, they are typically expected to exhibit human-like communicative behaviors, using speech and non-verbal expressions just as humans do. Gestures help in conveying information which speech alone cannot provide and need to be completed, as in referential, spatial or iconic information [HAB11]. Moreover, providing multiple modalities helps to dissolve ambiguity typical of unimodal communication and, as a consequence, to increase robustness of communication. In multimodal communication, gestures can make interaction with robots more effective. In fact, gestures and speech interact. They are linked in language production and perception, with their interaction contributing to an effective communication [WMK14]. In oral-based communication, human listeners have been shown to be well attentive to information conveyed via such non-verbal behaviors to better understand the acoustic message [GM99].

This topic can be addressed in the field of robotics where few approaches incorporate both speech and gesture analysis and synthesis [GBK06, SLO3], but also in the field of developing virtual conversational agents (talking avatars), where the challenge of generating speech and co-verbal gesture has already been tackled in various ways [NBM09, KW04, KBW08].

For virtual agents, most existing systems simplify the gesture-augmented communication by using lexicons of words and present the non-verbal behaviors in the form of pre-produced gestures [NBM09]. For humanoid robots the existing models of gesture synthesis mainly focus on the technical aspects of generating robotic motion that fulfills some communicative function, but they do not combine generated gestures with speech or just pre-recorded gestures that are not generated on-line but simply replayed during human-robot interaction.

Assignment
The goal of this thesis is to develop a gesture model for a credible communicative robot behavior during speech. The generation of gestures will be studied when the robot is a speaker and when it is a listener. In the context of this thesis, the robot will be replaced by an embodied virtual agent. This allows applying of the outcome of this work in both virtual and real world. It is possible to test the results of this work on a real robot by transferring the virtual agent behavior to the robot, when possible, but it is not an end in itself.

In this thesis, two main topics will be addressed: (1) the prediction of communication-related gesture realization and timing from speech, and (2) the generation of the appropriate gestures during speech synthesis. When the virtual agent is listening to a human interlocutor, the head movement is an important communicative gesture that may give the impression that the virtual agent understands what is said to it and that may make the interaction with the agent more effective. One challenge is to extract from speech, both acoustic and linguistic cues [KA04], to characterize the pronounced utterance and to predict the right gesture to generate (head posture, facial expressions and eye gaze [KCD14]). Synchronizing the gestures with the interlocutor speech is critical. In fact, any desynchronization may induce an ambiguity in the understanding of the reaction of the virtual agent. The gesture timing correlated with speech will be studied. In this work, generating the appropriate gesture during speech synthesis, mainly head posture, facial expressions and eye gaze, will be addressed.

Main activities
To achieve these goals, motion capture data during uttered speech will be acquired synchronously with the acoustic signal. Different contexts will be considered to achieve the collection of sufficiently rich data. This data will be used to identify suitable features to be integrated within the framework of machine learning techniques. As the data is multimodal (acoustic, visual, gestures), each component...
will be used efficiently in collecting complementary data. The speech signal will be used in the context of a speech-recognition system to extract the linguistic information, and acoustic features helps to extract non linguistic information, as F0 for instance. The correlation between gestures and speech signal will also be studied. The aim of the different analyses is to contribute to the understanding of the mechanism of oral communication combined with gestures and to develop a model that can predict the generation of gestures in the contexts of speaking and listening.

References


Skills

Good background in modeling, data analysis and machine learning. First experience in speech recognition or in using a deep learning technique will be appreciated.

Benefits package

- Subsidised catering service
- Partially-reimbursed public transport
- Social security
- Paid leave
- French courses

Remuneration

Gross Salary per month: 1982€ brut per month (year 1 & 2) and 2085€ brut/month (year 3)

Warning: you must enter your e-mail address in order to save your application to Inria. Applications must be submitted online on the Inria website. Processing of applications sent from other channels is not guaranteed.