Technological Trends in Natural User Interfaces

Ivan J. Tashev

Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA

0. Introduction

Each input or output modality of the human computer interface has roughly two parts: a sensor, and a recognizer. For example Microsoft Kinect brings two modalities: speech and gesture. We have a four element microphone array which with a speech enhancement pipeline forms the voice sensor. The recognizer in this case is the speech recognition. Let see some trends in the underlying technologies which enable progress in human machine interfaces.

1. Sensor fusion

The first trend is sensor fusion. While each of the underlying technologies is going to become better and better, this is a quantitative improvement and cannot be compared to the qualitative jump each one of these technologies caused when it appeared for the first time. We believe that there is a lot of potential in combining the data from different sensors to bring new quality. For example, in the first Kinect the listening beam was controlled by the sound source localizer, which pointed it to the loudest sound source, even if this is a vacuum cleaner. It is a much better solution to combine the outputs of the sound source localizer, the face detector, and the skeletal tracker, and to control the beam with the output of this sensor fusion block. At least we will point the beam towards a human face. And, with a little bit of studying of the dynamics of the human conversation, we will be able to predict who, from the multiple people in front of a Kinect device, is going to speak next and direct the listening beam in advance.

2. Better priors and context

The second trend is providing better priors and context to the human-machine interface. Subconsciously humans reflect the gender, the age, and the emotional status in the conversation with the other human. The technologies for estimation of the emotion from voice, face, and skeleton are already available. It is pretty straightforward to retrieve the gender and the age from the human voice. For example, let’s say that a voice query for a movie sounds something between “Die Hard” and “Sleepless in Seattle”. If it is requested by a 25 years old male, owner of the console, with a past record of ordering action movies, it is a safe assumption for the human-machine interfaces to believe that this is “Die Hard”. On the other hand if there is a second skeleton and a female voice in the room, it is most probably the more romantic “Sleepless in Seattle”.

3. Form HMIs to HsMIs

The third trend we see is the transition from Human-Machine Interfaces (HMIs) to Humans to Machine interfaces (HsMIs). This means converting the computer into an equal participant in the conversation between multiple humans. For example, if two people were talking about a nice dinner in an Italian restaurant they had last month, and then one of them asks the computer system for a nearby restaurant, a smart response would be: “There is a new Italian restaurant nearby with excellent reviews, would you like directions?” Besides the problems in building such a natural user interface, there are a lot of technological challenges. These include sound source separation, because humans tend to overlap each other when speaking, and building a “social” layer to detect the connections between different participants in the conversation since humans tend to talk about multiple topics at the same time.

4. Conclusion

With the proliferation of computing technologies providing new and better quality input modalities we can expect human-machine interfaces to closer and closer mimic the behavior of a real human.

Thank you!