Designing an Emotion Detection System for a Socially-Intelligent Human-Robot Interaction

Clément Chastagnol[1,2], Céline Clavel[1,2], Matthieu Courgeon[1], and Laurence Devillers[1,3]

[1] Département Communication Homme-Machine, LIMSI-CNRS, 91403 Orsay
[2] Département Informatique, Université Paris-Sud 11, 91403 Orsay
[3] Université Paris-Sorbonne 4

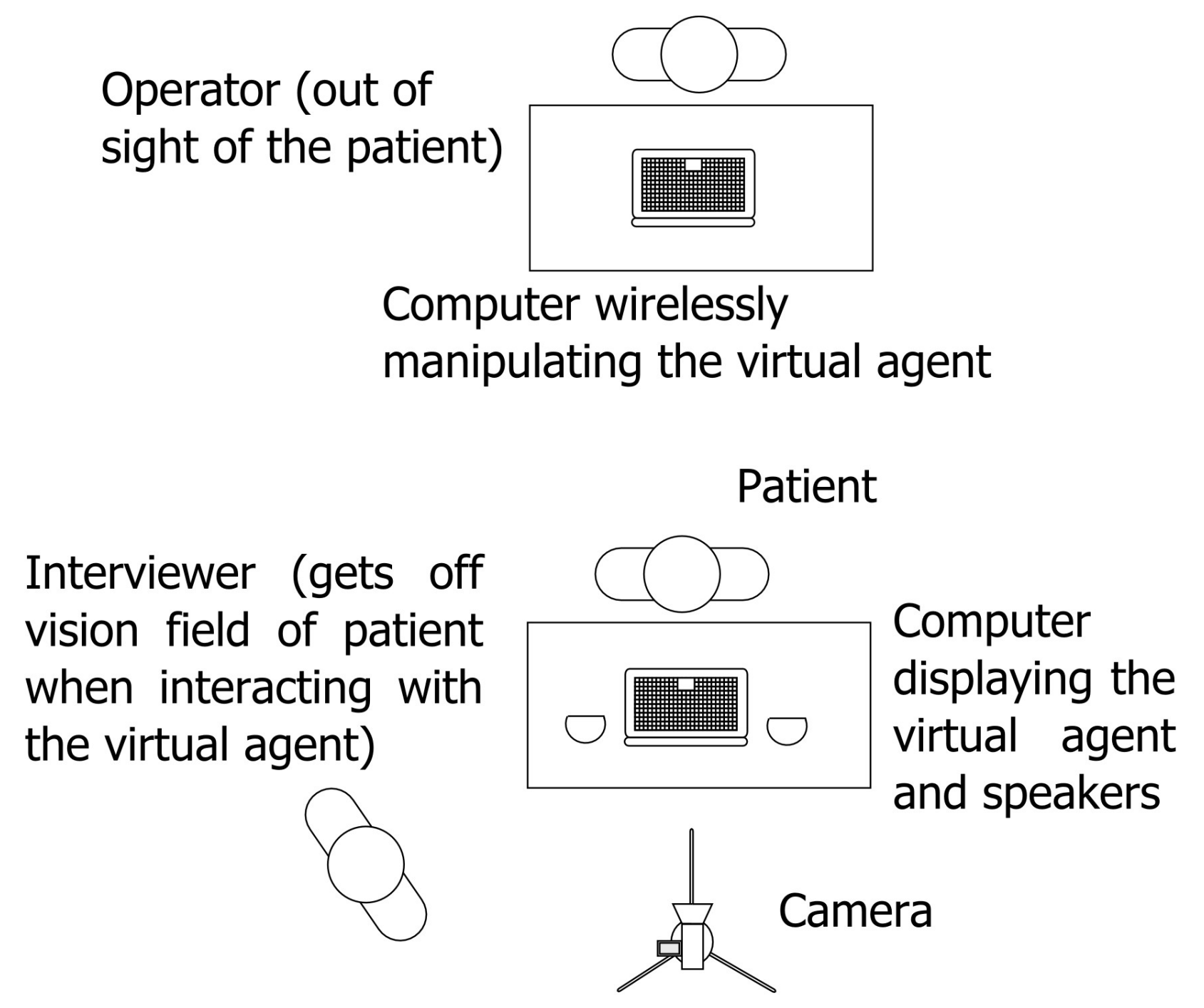
{cchastag, devil}@limsi.fr

Introduction

- French ARMEN project on assistive robotics: building a robot for elderly and disabled people.
- Collection of induced emotional data in interaction with a virtual character to build an emotion detection module, with the participation of 25 patients from medical centers.
- Specific difficulty: wide variety of voice (aged, degraded...) and emotional behaviors.
- Exploration of correlation between quantitative and qualitative interactional clues to build a measure of engagement.

Data collection

25 patients from medical centers, aged 25 to 91, interviewed on site in July 2011. Patients were interacting with a virtual character in a Wizard-of-Oz way. Four scenarii in daily life context, designed to induce emotions. Audio recorded in good quality + video.


Operator (out of sight of the patient)
Computer wirelessly manipulating the virtual agent
Patient
Interviewer (gets off vision field of patient when interacting with the virtual agent)
Computer displaying the virtual agent and speakers
Camera

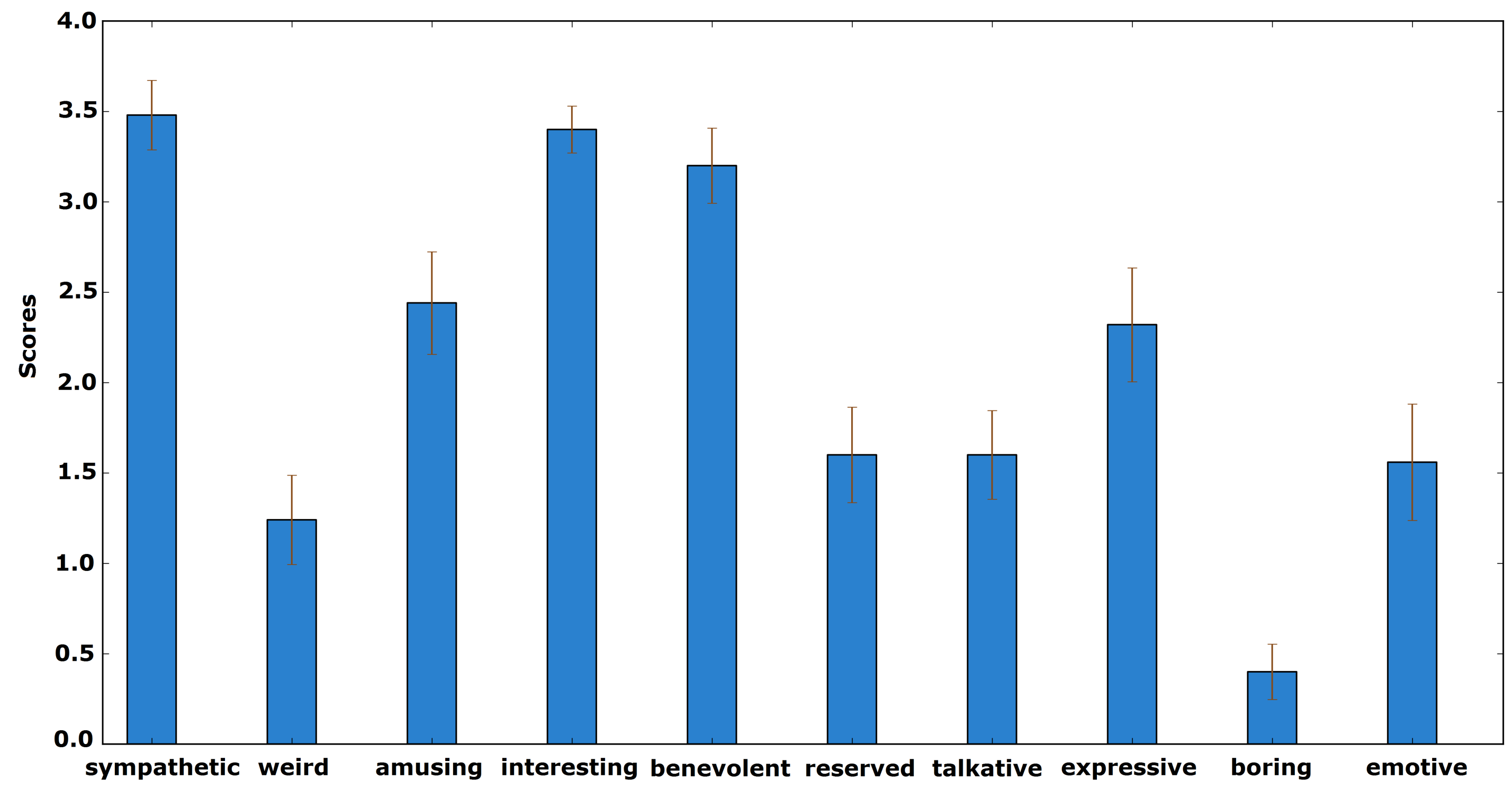
Three phases in the interview process:
- Introduction and explanations by the interviewer
- Interaction of the patient with the virtual character
- Answering to a questionnaire on the perception of the virtual character and the interaction

The scenarii were outlined by physicians and functional therapists, written by the authors and validated by the physicians. They are designed to satisfy several constraints:
- matching the test cases for the functionalities of the robot
- being close to the final user experience
- eliciting emotions to collect useful training data
- being easy to relate to for the subjects
- offering variability within a limited context to ensure robustness.

Additional information:
- 8.7 hours of audio and video data have been collected
- Audio data was manually labelled by two annotators on five coarse labels (Anger, Fear, Happiness, Neutral, Sadness) and a five levels Likert scale for Activation
- Agreement level for the emotion labels was around 63%

Questionnaire

The questionnaire helped us to understand two dimensions: how the users perceived the VC in terms of personality and how they perceived the interaction.



Virtual character traits attribution levels.

The virtual character was perceived positively by the subjects, with a high attribution of positive qualifiers and a low attribution of negative ones. The interaction was also deemed positive, with some differences between the subjects with respect to age.

First results for the emotion recognition module – Questionnaire analysis

Training data

Set name	ARMEN	JEMO
# of segments	545	1491
# of locutors	25	59
UAR score	45.1 %	68.0 %
# of Anger segments	92 (17%)	291 (19%)
# of Happiness segments	236 (43%)	359 (24%)
# of Neutral segments	136 (25%)	534 (36%)
# of Sadness segments	81 (15%)	307 (21%)

Details and results for the training sets.

We selected only the consensually annotated segments from the interaction phase. The Fear segments were not considered because there was too few of them. We compared the resulting training data to the JEMO corpus, featuring spontaneous and induced emotions collected in the frame of a game in lab conditions.

Experimental protocol

- Subsampling of the Neutral class, deletion of the Fear class
- Extraction of 384 acoustic parameters with the openEAR library (Interspeech 2009 Challenge configuration) (Schuller et al. 2010).
- Training of an RBF-kernel SVM with optimization of C and γ parameters; Leave One Speaker Out evaluation.
- UAR score.

Objective measures at the dialog level – Analysis of the questionnaire

Some objective measures of involvement in the interaction where derived from the segmented files:
- number of subject turns,
- number of answers to a question from the VC,
- speech duration,
- response time,
- number of overlapping segments

We analyzed the correlations of these measures with the questionnaire results (absolute value ≥ 0.4). We can note some interesting points:
- Subjects with a higher average response time found the VC to be less sympathetic (-0.471) and more bizarre (0.552).
- The age of subjects is negatively correlated with the VC being perceived as amusing (-0.464) and emotive (-0.494), and the interaction being perceived as captivating (-0.400) and entertaining (-0.433),
- but paradoxically, age is negatively correlated with the interaction being perceived as repetitive (-0.507).
- Subjects with a higher proportion of Happiness segments answered more quickly (0.460) and found the VC to be more communicative (0.413) and amusing (0.426).
- On the contrary, subjects with a higher proportion of Neutral segments have a higher average response time (0.506).
- There are almost no differences between genders, with exceptions for some perceived traits of the VC, which is judged less emotive (0.420), and the interaction less entertaining (0.514) by women than by men.

Conclusion and Discussion

- The data we collected with real end-users, sometimes with pathological and aged voices, is hard to work with compared to more classical databases.

- The analysis of interactional clues *wrt* a questionnaire on the perception of the virtual character by the users led to some interesting insights: the way subjects talk (response time, activation level...) is correlated to how they feel in the interaction. It also shows that it is important to take into account the differences *wrt* age between the speakers when designing a spoken dialog system.

- These first results will help us building the emotion detection module and working on a measure of engagement.

A few references

- Cassel, J. (2000). More Than Just Another Pretty Face: Embodied Conversational Interface Agents. In Communications of the ACM, volume 43, numéro 4, pages 70–78.
- Courgeon, M., Martin, J-C. et Jacquemin, C. (2008). MARC: a Multimodal Affective and Reactive Character. In Proceedings of the 1st Workshop on AFFective Interaction in Natural Environments, Chania, Crète.
- Delaborde, A., Tahon, M., Barras, C., and Devillers, L. (2009). A Wizard-of-Oz Game for Collecting Emotional Audio Data in a Children-Robot Interaction. In Proc. of the International Workshopon Affective-aware Virtual Agents and Social Robots, ICMI-MLMI, Boston, USA.
- Delaborde, A. et Devillers, L. (2010). Use of non-verbal speech cues in social interaction between human and robot: Emotional and interactional markers. In Proceedings of the 3rd ACM Workshop on Affective Interaction in Natural Environments, pages 75–80.
- Schuller, B., Steidl, S. et Batliner, A. (2009). The Interspeech 2009 Emotion Challenge. In Proc. of the 10th Annual Conference of the International Speech Communication Association, Brighton, UK.