International Journal of Humanoid Robotics
Vol. 15 (2018) 1850006 (31 pages)

DOI: 10.1142/S0219843618500068

Emotional Storytelling Using Virtual and Robotic Agents

Sandra Costa

Department of Industrial Electronics Engineering,
University of Minho, Minho, Portugal
scosta@dei.uminho.pt

Alberto Brunete*

Center for Automation and Robotics (CAR) UPM-CSIC,
Universidad Politecnica de Madrid, Spain
alberto.brunete@upm.es

Byung-Chull Bae

School of Games, Hongik University, Sejong, South Korea
byuc@hongik.ac.kr

Nikolaos Mavridis

Interactive Robots and Media Lab (IRML),
Athens, Greece
nmavridis@iit.demokritos.gr

Received 18 July 2016
Accepted 6 November 2017
Published 15 March 2018

In order to create effective storytelling agents three fundamental questions must be answered: first, is a physically embodied agent preferable to a virtual agent or a voice-only narration? Second, does a human voice have an advantage over a synthesized voice? Third, how should the emotional trajectory of the different characters in a story be related to a storyteller's facial expressions during storytelling time, and how does this correlate with the apparent emotions on the faces of the listeners? The results of two specially designed studies indicate that the physically embodied robot produces more narrative attention to the listener as compared to a virtual embodiment, that a human voice is preferable over the current state of the art of text-to-speech, and that there is a complex yet interesting relation between the emotion lines of the story, the facial expressions of the narrating agent, and the emotions of the listener, and that the empathizing of the listener is evident through its facial expressions. This work constitutes an important step towards emotional storytelling robots that can observe their listeners and adapt their style in order to maximize their effectiveness.

Keywords: Storytelling; robot and virtual agents; emotional affective response; eye blink analysis; facial expression analysis; non-verbal communication; posture analysis.

*Corresponding author.

1. Introduction

Storytelling is a form of oral communication with pre-historic beginnings, which serves as a means of acculturation as well as transmission of human history.[1] Traditionally, human storytelling has been one of the main means of conveying knowledge from generation to generation, but nowadays new technologies have also been used in this knowledge-sharing process.[2]

A way to view the process of storytelling is the following: First, the storyteller understands the narrative message that is conceived by the story author. Second, the storyteller delivers it to the listener in an effective way. Unlike written narrative communication, however, where the author communicates with the reader through the implicit communication channel (real author -> implied author -> narrator -> narratee -> implied reader -> real reader),[3] the storyteller performs storytelling face-to-face in real-time. Thus, the storyteller can infer whether the listener is paying attention to or being engaged in the story from the listener's responses or back channels such as verbal responses (e.g., acknowledgement tokens such as yeah, uh huh, or mm hm)[4] and non-verbal responses (e.g., head nodding, eye blinking, or smiles). When the negative backchannels (e.g., head down or blank expression from boredom throughout the storytelling) are continuously recognized, an effective storyteller will change his or her narration technique to keep the listener's engagement in the story. While the storyteller's narration techniques will be various depending on the listener profiles (e.g., age, education, preferences, etc.), emotional expression (either verbal or non-verbal) is a common quality of the effective storyteller. Recent work by Tang *et al.*[5] proposes a new human–robot communication framework using both verbal and non-verbal means.

The notion of attention, due to its importance in our everyday lives, has been studied in diverse disciplines including philosophy, psychology, cognitive science, neuroscience, and even computer science for the computational model of attention.[6] While this diversity has allowed elusive definitions on attention, a key concept of attention as mental awareness is that it is selectively done with some cognitive limits.[7] Attention is also addressed in Csikszentmihalyi's concept of "flow"[8] in which we can have ourselves completely engaged in our activities. In narrative, there have been also an attempt to give practical metrics for measuring narrative engagement, where four dimensions have been explored — narrative understanding, attentional focus, emotional engagement, and narrative presence.[9] Inspired by these previous studies, we mainly consider two narrative engagement factors — attentional focus and emotions in narrative.

The reader's emotional responses while reading a story in text can be either internal or external.[10] Examples of "internal" emotions (according to Ref. 10) are identification or empathy with the story characters which occurs when the reader enters the story world. Examples of "external" emotions include curiosity and surprise which occurs when the reader meets with the narrative discourse structure through text.[10] In the same vein, the storyteller can elicit the listener's emotional

responses either internally or externally. Specifically to arouse the listener's internal emotional responses, the storyteller can play an emotional surrogate of the characters or the narrator in the story. To enhance the listener's external emotional responses, the storyteller can pretend the listener's desirable emotional states (e.g., pretending curious or surprise more or less in an exaggerated way).

The storyteller agent or system can detect the listener's non-verbal responses using various sensor devices. The detection of positive back channels is an indication of the listener's satisfaction in the story. In this case the storyteller system will continue to tell the story with the current storytelling rhetoric. If some negative back channels over the specified threshold are detected, however, this might be a sign of the listener's disliking or inattentiveness.

We have mentioned the word "empathy" before, a word that has a very important role in this paper. According to Ref. 11, empathy is the ability to detect how another person is feeling, while Ref. 12 has defined empathy as: "Empathy accounts for the naturally occurring subjective experience of similarity between the feelings expressed by self and others without losing sight of whose feelings belong to whom". Empathy involves not only the affective experience of the other person's actual or inferred emotional state but also some minimal recognition and understanding of another's emotional state". Thus, empathy plays an important role in effective storytelling, as through observation of the listener's emotions, the storyteller can modulate his storytelling manner, in order to maximize effectiveness.

The storyteller has usually been a person, but recently some initial experiments with avatars as well as robots have taken place. If one wants to have humanoid robots in the role of learning companions and interaction partners to humans, engaging storytelling is a very important skill for them, according to Ref. 13.

Motivated by the above state of affairs, and towards our goal of effective emotional storytelling using Robotic and Virtual Agents, we performed two specially designed studies. In the first study we employed a virtual embodied conversational agent (Greta) as a storyteller that telling a story with emotional content expressed both by sound and facial expressions. In the second study we used a full-size highly-realistic humanoid robot (Aesop) with the same story material. In this paper we provide some initial empirical results of our studies, during which we observed the reactions of experimental subjects to artificial storytellers, through a combination of instruments: specially designed questionnaires, manual body language annotation, as well as automated facial expression and eye blink analysis.

1.1. *Research questions and expectations*

In this study, towards our ultimate purpose of effective emotional storytelling, we formulate three research questions (RQ1-RQ3) as follows:

- RQ1: Can a physical robot elicit more narrative attention when telling a story instead of a virtual agent? (choice of embodiment for storytelling agent).

- RQ2: While a story is told by the physical robot, will the participants show more non-verbal responses to a human voice recording as compared to a TTS voice recording? (choice of voice for storytelling agent).
- RQ3: Will the participants empathize with the emotions conveyed by the robot as story narrator? (effectiveness of affective performance of storytelling agent).

In order to answer RQ1 (effect of physical embodiment in listener's narrative attention), we measure narrative attention through the eye blink rate of the listener, as it is well known that through parasympathetic mechanisms blink rate correlates to attention. In other words, spontaneous blinking rates are different depending on the types of behavior (e.g., conversation > rest > reading[14]) or visual information[15]). We expect that the audience blink rate will be less for the case of the physical robot storyteller (Aesop), as compared to the virtual agent storyteller (Greta).

We furthermore expect the listener's narrative attention to vary throughout the story, and the climax of the narrative attention to be at the climax of the story. Thus, apart from our main question RQ1 which was related to the difference of listener attention across embodiments (robot vs. avatar), we also formulate a second sub-question — RQ1a: is there any difference between the various temporal parts of the story regarding blinking rate? RQ1 and RQ1a are addressed in experimental study 2 and study 1, respectively.

Regarding RQ2 (effect of choice of synthetic vs. real voice), we codify the participant's body language, expecting a real human voice to show signs of greater engagement as compared to the synthetic. This question is addressed in experimental study 2.

Finally, regarding RQ3, we compare empathy by measuring the similarity between the emotion line of the story and the emotion line of the observed facial expressions of the participants. In every story, we assume that there are different implied emotional lines over time for the characters, the narrator, and the listener. The listener's expected reaction is not the same as the narrators; rather, it is a function of all the above emotional lines. For example, the listener might feel anger about a sad character, because he or she might feel the situation is unfair. Our expectation is that the participants will empathize more with the story when the physical robot is the storyteller.

In the next section relevant background literature using robots or virtual agents in storytelling scenarios is presented. Section 3 presents the architecture of our system, and Secs. 4 and 5 feature the procedures and results in the experiments of Studies 1 and 2. Section 6 contains the discussion of the results, and conclusions are presented in Sec. 7.

2. Background

2.1. *Storytelling using virtual agents*

Several studies on emotionally expressive storytelling have been conducted using virtual agents. However, none have used automated analysis of non-verbal behavior

of the listeners with systems such as SHORE[a] and FaceAPI,[b] as we do in this paper. For example, Silva *et al.*[16,17] presented Papous, a virtual storyteller using a synthetic 3D person model with affective speech and affective facial/body expression. Papous could express six basic emotions (joy, sadness, anger, disgust, surprise, and fear) with different emotional intensity in the input text. The emotion tagging in the input script was simply made just for the narrator (i.e., storyteller) without considering possible emotions from different story characters or the listeners.

In the Conveying Affectiveness in Leading-edge Living Adaptive Systems (CALLAS), an FP6 European research project, an integrated affective and interactive narrative system using a virtual character (Greta) was presented.[18] The proposed narrative system could generate emotionally expressive animation using a virtual character (Greta) and could adapt a given narrative based on the detected user emotions. In their approaches, the listeners' emotional states could be detected through emotional speech detection, which were applied to their interactive narrative system as either positive or negative feedback.

As a part of CALLAS project, Bleackley *et al.*[19] investigated whether the use of an empathic virtual storyteller could affect the user's emotional states. In their study, each study participant was listened to a broadcast news about earthquake disaster twice — first, with only the voice along with relevant text, image, and some music; next, with an empathic virtual storyteller (Greta) along with the same conditions as the first. The Self Assessment Mannequin (SAM) test was used to measure the possible change of the user's emotional states in terms of Pleasantness, Arousal, and Dominance (PAD) level before and after listening to the news story. The results showed that the use of empathic virtual storyteller influenced the user's emotional states — the participants' average level of pleasantness and dominance were decreased respectively when the Greta was used as a proxy of empathic virtual storyteller. Our study on virtual storyteller was inspired by this mock-up study but we employed a fictional story with multiple characters and various narrative emotions (e.g. happiness, sadness, surprise, etc.) in it, and furthermore, we employ automated as well as manual analysis of a both face and whole-body non-verbal behavior. The former controls the story logic (such as story flow and coherency) and the latter keeps track of the reader's anticipated emotions. The notion of tracing the reader's anticipated emotions has some superficial similarity with our approach, but ours is focused rather on recognition of the listener's narrative attention and storyteller's empathizing with the emotions of the story characters.

In order to measure the listener's narrative attention in an objective way, various factors (e.g., glancing, standing, nodding, or smiling) can be used to evaluate attentive engagement of the listener using visual information.[3] In our study we included eye blinking as a measurement of the listener's narrative attention level based on the empirical studies claiming that inhibition of blinking is closely related to

[a] http://www.iis.fraunhofer.de/en/bf/bsy/produkte/shore/.
[b] http://www.faceapi.com/.

the intent of not losing important visual information.[14,15] For example, according to Bentivoglio *et al.*,[14] eye blink rate shows a tendency of decreasing while reading (which requires more attention) and increasing during conversation (which requires less attention).

2.2. *Storytelling using robots*

Personal Electronic Teller of Stories (PETS) is a prototype storytelling robot to be used with children in rehabilitation.[20] This robot was used remotely by children using a variety of body sensors adapted to their disability or rehabilitation goal. The children were meant to teach the robot how to act out different emotions such as sadness, or happiness, and afterwards to use storytelling software to include those emotions in the stories they wrote. The authors believed this technology was a strong encouragement for the children to recover quickly and may also help the children learn new skills. PETS' authors focused on guidelines for cooperation between adult and children, and design of game scenarios. The use of robots as storytellers for educational purposes has been increasingly researched, as for example Wong *et al.*[21] or Ying *et al.*[22]

The ASIMO robot was used in a storytelling study where the goal was to verify how human gaze can be modelled and implemented on a humanoid robot to create a natural, human like behavior for storytelling. The experiment performed in Ref. 23 provides an evaluation of the gaze algorithm, motivated by results in the literature on human-human communication suggesting that more frequent gaze toward the listener should have a positive effect. The authors manipulated the frequency of ASIMO's gaze between two participants and used pre and post questionnaires to determine the participants' evaluation of the robot's performance. The results indicated that the participants who were given more frequent gaze from ASIMO performed better in a recall task.[23]

The GENTORO system used a robot and a hand held projector for supporting children's story creation and expression. Story creation included a script design, its visual and auditory rendering, and story expression as a performance of the script. The primary goal of the presented study was to clarify the effects of the system's features and to explore its possibilities. Using post-experimental questionnaires answered by the children, the authors affirmed that children had considerable interest in the robot, because it behaved like a living thing and always followed a path on a moving projected image.[24]

Pleo, a robotic dinosaur toy, has been used to mix physical and digital environments to create stories, which were later programmed with the goal of controlling robotic characters. Children created their stories, and programmed how the robotic character should respond to props and to physical touch. The system gave children the opportunity and control to drive their own interactive characters, and the authors affirmed they contributed to the design of multimodal tools for children's creative storytelling creation.[25]

In Ref. 26, Lego Mindstorms robotics kits were used with children to demonstrate that robots could be a useful tool for interdisciplinary projects. The children constructed and programmed the Lego robots, addressing the dramatisation of popular tales as the final goal. The study results showed the applicability of robots as an educational tool, using storytelling as a background, developing thinking, interaction and autonomy in the learning process.

The results of two-year's research in a classroom of children with intellectual disabilities and/or autism are described in Ref. 27. The PaPeRo robot was used to enhance storytelling activities. The authors found that the length of stories produced by the children continually increased and the participants began to tell more grammatically complex stories.

From Ref. 2, a survey on storytelling with robots and other simple projects are presented. Summarizing, the main users or the target group of the presented studies are normally developed[24] or disabled children,[20] either with the goal of teaching/learning or rehabilitation. Adults were involved in some projects,[23] but usually as teachers.[26] The robots used in the previous projects were mostly small mobile robots (e.g., Lego Mindstorms or Pleo), and the outcomes of the studies were mainly design, pedagogics, and prototypes in authoring, learning or mixed environments.

Comparing our work with the studies presented above, in the study presented in this paper, the target group is composed only of adults. However, our methodology could also be used with children. So, we have a different and more general listener, and most importantly we are directing our goals towards the automated observation of the emotional and non-verbal communication provided by the participants while reacting to the story told by the robot or the virtual agent. One of the main novelties of our study thus consists in the analysis of the whole-body non-verbal communication, and the automated analysis of the facial expressions made by the participants during the storytelling experiments, as well as in the direct comparison between virtual avatar and real android robot embodiments. For example, robot embodiments have been used by Kennedy *et al.*[28] for children education.

3. System Design

In this section the architecture of the system is described. The system employs verbal and non-verbal rhetoric (e.g., emotional speech and facial expression) as a way of empathizing storytelling. The listener's narrative attention and his/her emotions while listening to the story is recognized using special automated software analyzers (FaceAPI and SHORE), in conjunction with standardized human observation and formalized description (for the case of non-facial body language).

3.1. *Overall architecture*

Two embodiments of the storytelling agent are utilized in this paper: a virtual agent (Greta[29]), and a physical humanoid robot (Aesop, a version of the Ibn Sina robot described in Refs. 30–32).

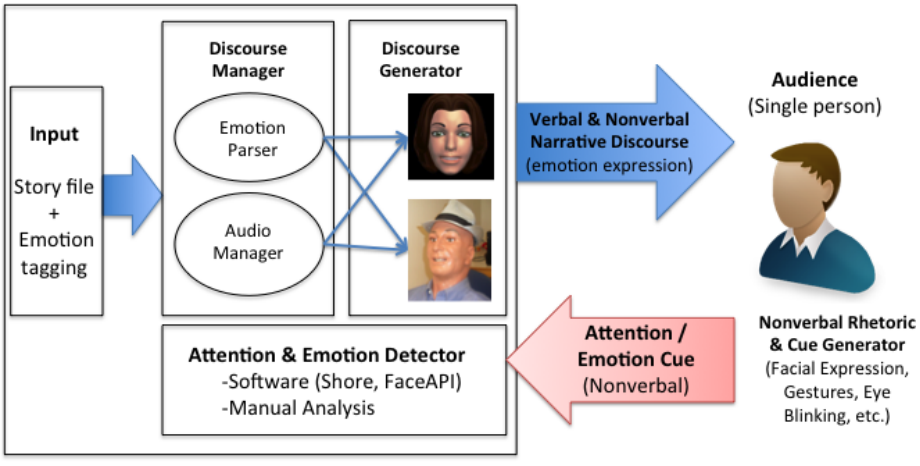


Fig. 1. System architecture.

As illustrated in Fig. 1, our proposed system consists of three main components: Discourse Generator, Discourse Manager and Attention and Emotion Detector. The input text story file was manually tagged with possible emotions and was formatted using FML–APML (Function Markup Language — Affective Presentation Markup Language).[33] FML–APML is an XML-based markup language for representing the agent's communicative intentions and the text to be uttered by the agent. This version encompasses the tags regarding emotional states which were used to display different emotions both on the virtual and on the robotic agents.

3.2. *Discourse manager*

Discourse Manager consists of two modules: Emotion Parser and Audio Manager. The Emotion Parser extracts emotion information from the input text file and formats it to be used by Greta or Aesop. The Audio Manager selects if the audio will be produced by a TTS systems or is provided as a pre-recorded voice.

3.3. *Discourse generator*

This module creates the storytelling experience. There are four options depending on the selected agent (Greta or Aesop) and the audio generation (TTS or pre-recorded voice). For Greta, if TTS is chosen, both video (with facial expressions corresponding to emotions) and audio and generated by a computer. If pre-recorded voice is chosen, the video (with facial expressions) is merged with the provided audio file.

The Aesop physical humanoid robot, on the other hand, uses its facial expression capabilities to show the emotions provided by the emotion parser, while the voice comes from TTS or pre-recorded audio.

3.4. *Attention and emotion detector*

This module takes as input a video from a listener-observing camera in real-time and analyzes the listener response in terms of emotions and narrative attention. In this paper Sophisticated High-speed Object Recognition Engine (SHORE)[34] is used to recognize the listener emotional facial expressions, and FaceAPI,[35] a real-time face tracking toolkit, to extract narrative attention related features (eye-blink) from the listener's face.

SHORE[34] has a face detection rate of 91.5%, and the processing speed of full analysis including facial expressions is 45.5 fps. SHORE recognizes the following facial expressions: Happy, Surprised, Angry, and Sad. The software is capable of tracking and analyzing more than one face at a time in real-time with a very high robustness especially with respect to adverse lighting conditions.

FaceAPI[35] provides an image-processing modules for tracking and analyzing faces and facial features. FaceAPI provides real-time, automatic monocular 3D face tracking, and it tracks head-position in 3D providing X, Y, Z position and orientation coordinates per frame of video. FaceAPI also enables blink detection, and can also track 3 points on each eyebrow and 8 points around the lips.

4. Experimental Study 1

This study mainly addressed experimental questions RQ2, dealing with the effect of voice choice on listener engagement, and RQ3, focusing on the relation between the emotional line of the story and the emotional reactions of the listener.

4.1. *Procedures*

We adopted a short story titled "The cracked pot" (consisting of 12 sentences and about 250 words) as a story material. This story is based on a Chinese parable, and has a lot of attractive features for our study. First, it has the right length - neither too long, nor too short: a little more than two minutes (2:09 in our narration). Second, it has a main character (the cracked pot), which we expect that the listener will feel affection for. Third, it has an active emotional trajectory of intermediate complexity.

In "The cracked pot" a heterodiegetic narrator (i.e., the narrator who is not present in the story world as a character) narrates a story about three characters (a woman, a perfect pot, and a cracked pot) with omniscient point of view.

In order to annotate the emotion line of the story, five adults individually tagged the possible character emotions sentence by sentence. The emotion category was limited to six basic emotions (Happiness, Sadness, Anger, Fear, Disgust, and Surprise) with intensity range from 0 (Not at all) to 10 (Extreme). The tagged data were collected and averaged with the confidence ratio based on the number of the responses (Resulting tagged emotions in Table 1).

Figure 2 shows the emotion line dynamics obtained from the emotion tagging by our human annotators in which the intensity of each emotion was obtained

Table 1. Story material (The Cracked Pot) and tagged emotion intensity (H for Happiness, S for Sadness, Su for Surprise).

Sentence	Narrator
An elderly Chinese woman had two large pots, each hung on the ends of a pole, which she carried across her neck (0–10 s).	H 1(0.2)
One of the pots had a crack in it while the other pot was perfect and always delivered a full portion of water (10–22 s).	S 4(0.4) H 2(0.2)
At the end of the long walk from the stream to the house, the cracked pot arrived only **half full** (22–30 s).	S 4.5(0.8)
For a full two years this went on daily, with the woman bringing home only one and a half pots of water (30–39 s).	S 5.5(0.2)
Of course, the perfect pot was proud of its accomplishments (39–45 s).	H 4.5(0.4)
But the poor cracked pot was ashamed of its own imperfection, and miserable that it could only do half of what it had been made to do (45–53 s).	S 8.5(0.4)
After 2 years of what it perceived to be bitter failure, it spoke to the woman one day by the stream (53–62 s).	S 6(0.6) Su 1(0.2)
"I am ashamed of myself, because this crack in my side causes water to leak out all the way back to your house (62–75 s)."	S 7(0.4)
The old woman smiled, "Did you notice that there are **flowers** on your side of the path, but not on the other pot's side?" (75–91 s)	Su 4.5(0.8)H 5.5(0.4)
"That's because I have always known about your flaw, so I planted flower seeds on your side of the path, and every day while we walk back, you water them." (91–106 s)	H 5.4(1.0) Su 5.5(0.4)
"For two years I have been able to pick these beautiful flowers to decorate the table (106–113 s).	H 7(0.6)
Without you being just the way you are, there would not be this beauty to grace the house (113–125 s)."	H 7.4(1.0)

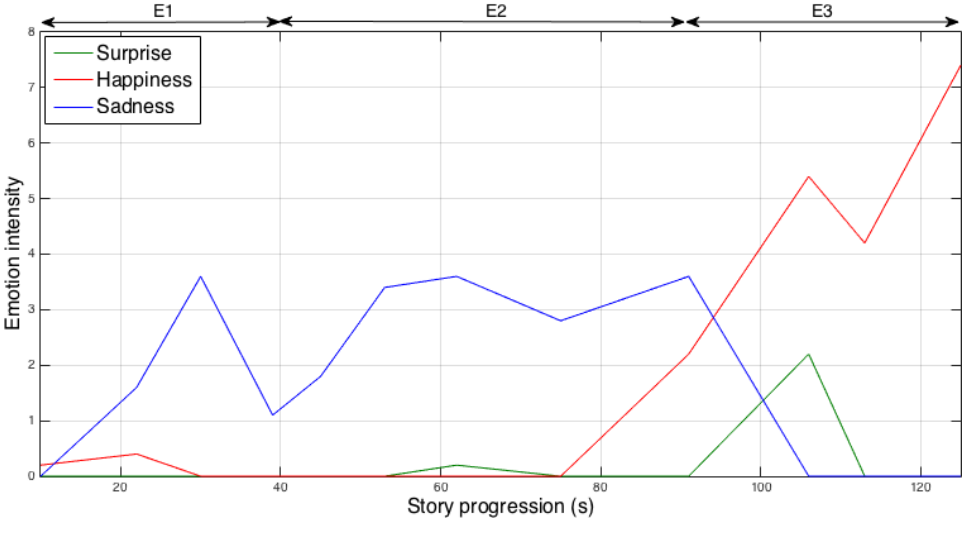


Fig. 2. The narrator's emotion dynamics in the cracked pot story (based on the tagged emotions).

considering the confidence ratio. For example, as seen in Fig. 2, the emotion of surprise was present as a transition emotion from a negative emotional state (i.e., sadness) to a positive emotional state (i.e., happiness).

4.1.1. *Experimental design and participants*

We designed a 2×2 between subjects with two factors (presentation type; gender) and two levels each (audio only and audio with video; male and female). A total of 20 participants (10 women, 10 men), who were students, staffs, and researchers from New York University Abu Dhabi, were volunteer participants in the experimental study 1. Their ages ranged from 18 to 60 years old. Each participant was arbitrarily assigned, while balancing the gender ratio, to one of the two groups. Ninety percent of the participants used English as a foreign language, while the others were native speakers. The participants in one group (Group A: audio only) listened (individually) to only the pre-recorded audio story which is narrated by a human storyteller (an amateur female voice actor; no embodiment whatsoever); the participants in the other group (Group B: audio with video) listened (individually) to the same audio story with video in which Greta expressed her emotions using facial expressions.

4.1.2. *Experimental setup*

The cracked pot story was recorded using an amateur female voice actor.

When participants entered the room where the study was conducted, they were guided to sit on a chair. A video camera was set up in front of the participants. The participants were given a Self Assessment Manikin (SAM) test sheet to describe their current emotional states and answered a pre-experiment questionnaire consisting of the questions about their demographic information. The purpose of a SAM test utilization was to check if the affective states of the participants are normal before and after the study. If someone is having an extremely unstable affective states before the study, for example, the study participant's data should be discarded.[36] As a criterion for the extreme affective state, we checked if the participants were feeling either extremely pleasant/unpleasant (that is, max/min values on the Pleasant dimension) or extremely aroused/calm (that is, max/min values on the Arousal dimension).

The participants in Group A listened to the audio story through the speaker in the room, without any other relevant material to the experiment; the participants in Group B watched a 50-inch TV screen on the wall in which Greta showed her emotional facial expressions according to the same audio story. Facial expressions of participants were recorded under their agreement. After the storytelling is over, the participants were asked to provide ratings on a 7-point scale, ranging from "not at all" (1) to "very much" (7) about their story appreciation. They were also asked to provide their opinions about the experiment as open questions. Finally they described their current emotional states using the same SAM test.

The questions set in the experiment questionnaire included the following 8 questions, where Q1 and Q4 are for the participants' engagement in the story;

Q2 and Q5 for the participants' empathizing with the main character; Q3 and Q6 for the participants' liking for the story; and Q7 and Q8 for the investigation of the possible differences between the emotions that the participants feel and the emotions that they notice during the story progression.

Q1. How interesting was the story?
Q2. How sorry did you feel for the bad pot in the beginning of the story?
Q3. How much did you enjoy listening to the story?
Q4. How much did you want to know how the story would end?
Q5. How happy did you feel for the bad pot at the end of the story?
Q6. How much did you like the story?
Q7. What emotions did you feel while listening to the story? (Please describe all the emotions you felt.)
Q8. What emotions did you notice while listening to the story? (Please describe all the emotions you noticed.)

4.1.3. *Evaluation tools*

Besides standard statistical software to analyze the data produced from the experiments, SHORE[34] and FaceAPI[35] were used. SHORE was used for the recognition of the listener's facial expression; FaceAPI for eye-blinking detection.

In particular for this study, it is important the recognition of facial expressions (happy, surprised, angry, and sad) and the detection of in-plane rotated faces (up to +/− 60 degree), in order to be able to examine the facial expressions displayed by the participants and their relation with the ones conveyed by the story, towards research question RQ3. Classification accuracy of basic emotions through facial expressions typically ranges between 85% and 95%.[37]

4.2. *Results*

None of the participants had to be excluded due to their performance in the SAM test. The summary data including the participants' average ratings and their standard deviations are shown in Table 2 through Table 4 in terms of the three story appreciation factors - liking, engagement, and empathizing.

A two way ANOVA with replication (in which two factors are presentation type and gender) for the three narrative engagement elements (liking, engagement, empathizing) showed no significant differences in main effects and interactions. Some possibly interesting gender difference, however, has been found in empathizing

Table 2. Comparison of the mean ratings for Liking (Q3+Q6) between two groups and gender in 7-point scale rating - M(SD).

Liking (Q3+Q6)	All participants	Male	Female
Audio only	4.45 (1.90)	**3.7 (1.77)**	**5.2 (1.81)**
Audio with video (Greta)	4.4 (1.67)	4.2 (1.48)	4.6 (1.67)

Table 3. Comparison of the mean ratings for Engagement (Q1+Q4) between two groups and gender in 7-point scale rating - M(SD).

Engagement (Q1+Q4)	All participants	Male	Female
Audio only	4.6 (1.76)	**3.9 (1.73)**	**5.3 (1.57)**
Audio with video (Greta)	4.85 (1.57)	4.8 (1.93)	4.9 (1.20)

Table 4. Comparison of the mean ratings for Empathizing (Q2+Q5) between two groups and gender in 7-point scale rating - M(SD).

Empathizing (Q2+Q5)	All participants	Male	Female
Audio only	4.25 (2.22)	**3.3 (2.06)**	**5.2 (2.04)**
Audio with video (Greta)	4.7 (1.45)	4.6 (1.58)	4.8 (1.40)

questions (Q2 and Q5), where $F(39, 77) = 0.643$; $p = 0.072$. Overall, the differences between the male and the female participants were greater in audio only presentation type than in audio with video presentation type. Overall, the study data indicate that the female participants might be relatively less affected by visual stimuli than the male participants, though it is not statistically supported in this study.

In addition to the analysis of the questionnaire from the participants, we also analyzed the facial expressions of them and compared their emotion trajectory with that of the narrator. Figure 3 shows the emotional analysis of the listener, which was made through automated classification of facial expressions using the SHORE

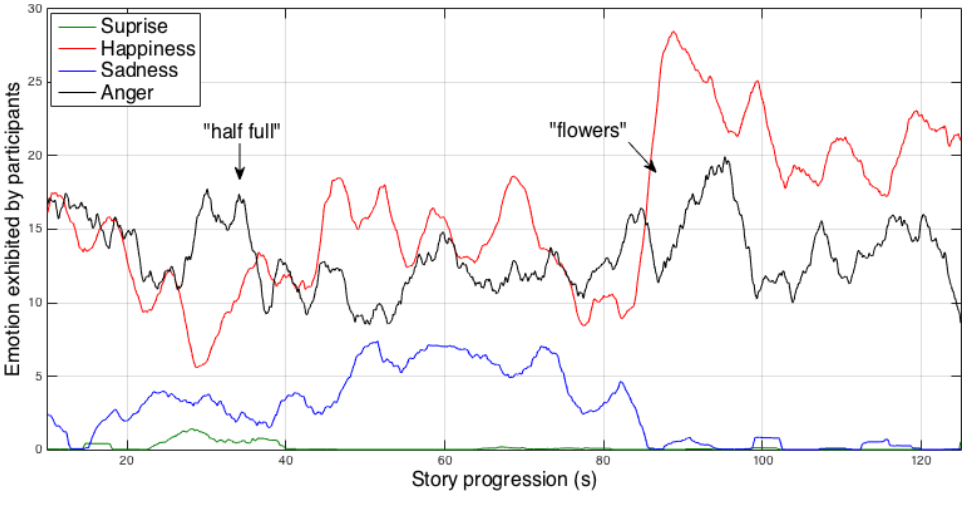


Fig. 3. The participants' emotion dynamics in the cracked pot story (based on the facial expressions analyzed by FaceAPI).

software. It shows the different emotions measured from the participants' faces while watching the video. A very interesting comparison here is with Fig. 2, which contains the facial expressions that the storytelling agent (robot or avatar) was programmed to perform during storytelling, on the basis of the tagged emotions of Table 1. Let us have a deeper look.

First, note the overall synchronization of the main transitions, between Figs. 2 and 3. In Fig. 3, two keywords have been overlapped to facilitate the analysis. The storytelling time-line of Fig. 2 contains a number of important events E1-E3, on the basis of the observed emotional transitions:

- E1: There is a sharp narrator sadness peak (blue line) that occurs around $t = 30\,\text{s}$ (almost in sync with the words "half full" in the story text), with a roughly triangular supporting ramp lasting between $t = 23\,\text{s}$ and $t = 37\,\text{s}$.
- E2: There is a second more stable narrator sadness plateau lasting roughly between $t = 45$ and dominating till $t = 90$ or so.
- E3: Happiness takes over the narrator's facial expressions from $t = 90$ approx. (almost in sync with the word "flowers" in the story text) until the end of the story.

Correspondingly, moving from the narrator emotional facial expressions in Fig. 2 to the resulting listener emotions as witnessed by automatically analyzed facial expressions in Fig. 3, one can notice the following events:

- E1': In good synchrony to E1, the sharp narrator sadness peak (blue line) produces a marked decrease in happiness and increase in anger in the listener ($t = 23\,\text{s}$ to $t = 37\,\text{s}$)
- E2': During the story period where the hero of the story (cracked pot) is sad and no positive signs appear on the horizon (and strong words such as "miserable", "bitter failure" are heard, the listener experiences increasing and then sustained sadness too, which is also the emotion the narrator expresses in E2 ($t = 45\,\text{s}$ to $t = 90\,\text{s}$)
- E3': Roughly when the word "flowers" is heard ("The old woman smiled, did you notice that there are flowers on your side of the path?"), there is a great increase in apparent happiness in the listener which after the first peak is sustained all the way to the end of the story, in response to E3 ($t = 90\,\text{s}$ until $t = 125\,\text{s}$)

Thus, what is apparent is that:

- There exists synchrony between story content, narrator facial expressions, and resulting listener facial expressions.
- The relation between these three time-lines is not a simple equality or one-to-one relation, but contains both its own "harmonies" as well as "dynamics". By "harmonies" we are referring to the relations of the emotions across the implicated time-lines. For example, during event E3, the happiness of the narrator is connected to the happiness of the listener in E3', and this is a simple equality relation (Narrator Happy - listener Happy). However, this is not the case in E1: there, the sadness of the narrator is reflected to anger in the listener; but this relation does

not always hold: for example, the narrator sadness in E2 causes an increase in listener sadness in E2', and not anger as it did in E1. We will further discuss these important observations below.

- The baselines (average value and scale) in the four emotional lines of Fig. 3 are different for each emotional component, and thus relative changes might well be a stronger indicator rather than absolute values. A more detailed discussion and elaboration will follow in the next sections.

Emotion can be modelled via various theories and models of emotion.[38] Among them a two-dimensional emotion model (or the circumplex model of affect[39]) is often employed, where various types of emotion are represented by two dimensions — arousal and valence. The former is about the intensity of emotion; the latter about the pleasantness of emotion. In order to better understand the results, we also analyzed the possible differences between the average valence values of the narrator and those of the participants over the story time-line. We have considered the formula in Eq. (1), where h stands for happiness, su for surprise, a for anger and s for sadness. m is the mean of each vector. Overall it shows the relation between positive feelings (that is, happiness and surprise) and negative feelings (sadness and anger). It could be argued whether surprise is either positive or negative since it can be a transition emotion either to negative valence or to positive valence. We included it, however, as a positive emotion under the context of the cracked pot story where surprise contributes to the addition of positive valence (that is, transition from sad emotion to happy emotion).

$$v[t] = (h[t] - m_h + su[t] - m_s - (a[t] - m_a + s[t] - m_s)). \tag{1}$$

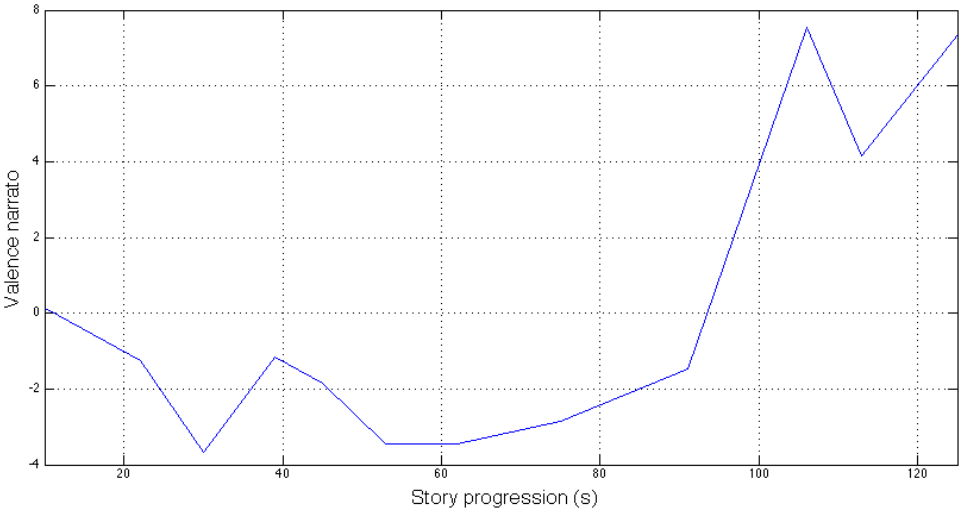


Fig. 4. The narrator's average valence values based on the formula in Eq. (1), where negative valence values indicate negative feelings; positive valence values indicate positive feelings.

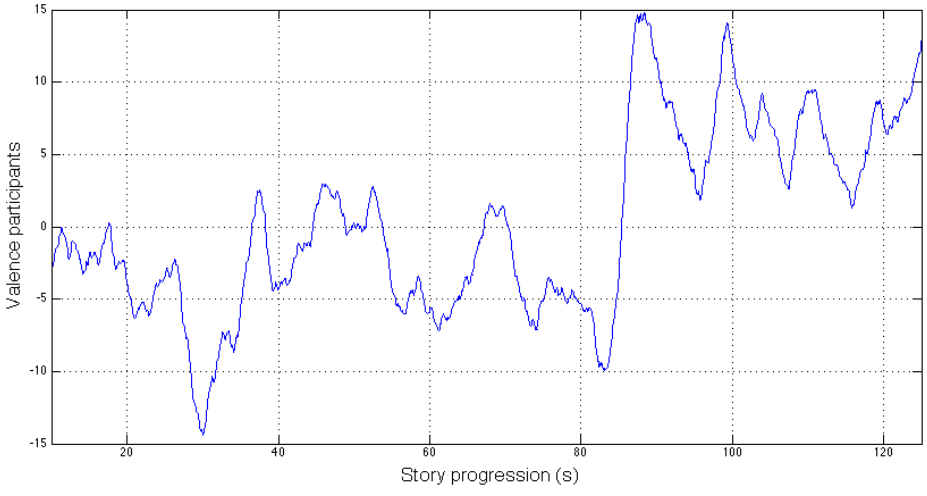


Fig. 5. The participants' average valence values based on the formula in Eq. (1), where negative valence values indicate negative feelings; positive valence values indicate positive feelings.

Figures 4 and 5 show the results of the narrator and the participants, respectively. It is possible to see that they have a similar pattern, meaning that in terms of valence (positive and negative feelings), the participant empathizes with the narrator.

Now let us move on from facial expressions to eye blink rate. Figure 6 shows the average blinking frequency of the listener during storytelling. The story has been

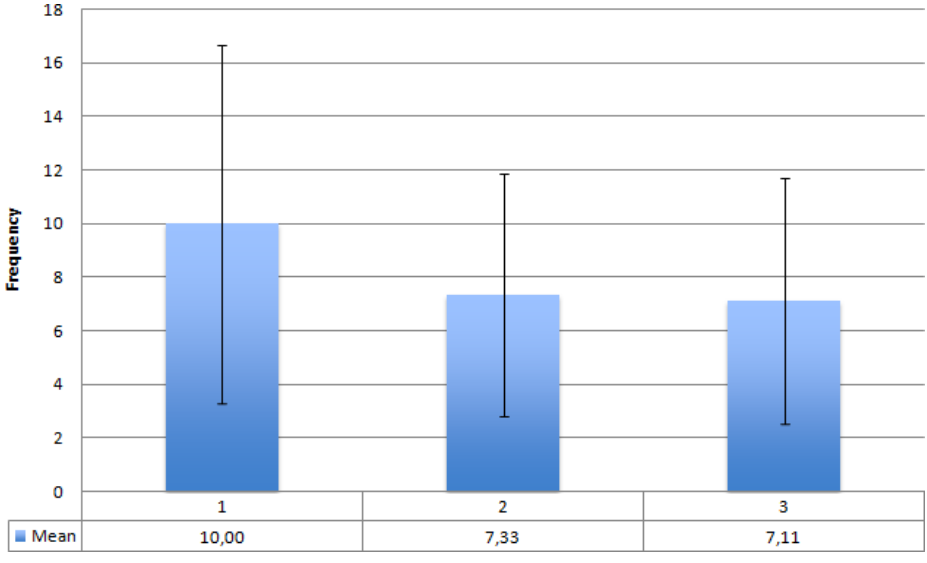


Fig. 6. Average number of eye blinking of the participants in subject group B during storytelling (error bars show the standard deviation).

divided in three parts to compare the number of eye blinks in each of them. It can be seen that as the story progresses the number of eye blinks decreases, signaling an increase in participants' narrative attention as they got involved in the story. A t-test performed over this data shows that the differences between eye blink rates between the story sections are indeed statistically significant ($p < 0.05$), providing evidence towards question RQ1a: i.e., our empirical data supports the hypothesis that as we are reaching towards the climax which occurs almost at the end of the story, apparent listener attention is increasing.

5. Experimental Study 2

As it was mentioned previously we wanted to verify the differences between a virtual and a robotic agent in a storytelling scenario, thus addressing research question RQ1. A new study with different participants was thus designed. In addition to this, special attention was now given also to the participants' non-facial body language, which was video-taped and annotated using a special formal scheme (used by the Observer XT 11 software by Noldus[40]) that will be described.

5.1. *Procedures*

Some elements used in Study 1 were used in Study 2, which are identified in the corresponding section, and the similarities and differences are pointed out. The same story as used in Study 1 was employed.

5.2. *Participants*

A total of 40 (8F, 32M) students and staff of NCSR-D Research Center (National Center for Scientific Research Demokritos) volunteered, with ages from 20 to 55 years (21 - 20/29 years old (YO); 14 - 30/39 YO; 4 - 40/49 YO; 1 - 50+). Each participant was arbitrarily assigned to one of the two groups. All used English as a foreign language.

5.3. *Experimental setup*

The participants in Group A listened to the story with a human voice recording and the participants in the Group B listened to the same audio story with a voice produced by text-to speech software. Besides having Greta telling the story to the participants, Aesop, a humanoid robot told "The cracked pot" story. This robot was already used in other scenarios[41,30] but in this study it was moving only its head, and not its hands, displaying different facial expressions according to the emotion conveyed in the story.

In the experiments, the participant was welcomed into a room and he/she was requested to sit on a chair. The room setup includes a table, the robot at one side and the participant at the other, and two cameras on both sides of the robot. The participant was requested to fill in a consent form, and an online questionnaire,

the same as used in Study 1. The SAM test was applied before and after the experiment, with the same goal as in Study 1.

All participants listened to the story in which Greta showed her emotional facial expressions accordingly. Half of the same participants also observed the Aesop robot narrating the story using a human voice and facial expressions, and the other half using a TTS voice and facial expressions. The order in which the participants listened to the story (either by Greta or by Aesop) was random.

When the storytelling was over, the participants were asked to provide ratings on a 7-point Likert scale ranging from "not at all" (1) to "very much" (7) about their story appreciation. Open questions were also asked to get opinions about the experiment.

5.3.1. *Evaluation tools*

Besides the same evaluation tools used in Study 1, the videos produced during the experiments were examined using the Observer XT 11 program by Noldus.[40] This software is a video annotation tool, and it was used to code events related with non-verbal body communication. Three different categories of behaviors were defined to be identified in the videos: hand and arm gestures, legs gestures, and head position. Inside these categories, the codes used were:

- Hand and Arm Gestures: steeping hands, evaluation, index finger, chin stroking, crossed arms;
- Legs Gestures: crossed legs, 4 leg lock, leg clamp;
- Head Positions: neutral head position, interested position, disapproval position, hands behind head.

These events were chosen from Ref. 42 having in mind participants' behaviors which would help us to answer RQ2 and RQ3. Summarily, some of these behaviors might help us to evaluate the non-verbal communication performed by the participants, indicating for example that the participants are interested, listening, evaluating, or disagreeing with the situation they are involved in.

Observation of gesture and congruence of the verbal and non-verbal channels are the keys to accurate interpretation of body language. However, all gestures should be considered in the context in which they occur.[42] Head orientation was suggested by Ekman and Friesen to be an indicator of gross affective state (positive/negative) as well as intensity of emotion.[43] A user study using the PAD model to assess the perception of affect from head motion during affective speech reported that head motion corresponding to distinct affective states is described by different motion activation, range, and velocity.[44]

Several studies regarding hand and arm movements show significant results for distinguishing between affective states.[45,46] These are recognized above chance level in full-light videos of hand and arm movements,[47,48] animated anthropomorphic and non-anthropomorphic hand models displaying abstract movements.[49]

To ensure inter-rater reliability 10% of the videos were re-coded by a second independent coder (Cohen's kappa $k = 0.64$). This is acceptable, as having a Cohen's kappa value higher than 0.60 suggests a good agreement between the raters.[50] When the coders were analyzing the behavior of the listener, they observe the videos from both cameras simultaneously and were able to hear the whole interaction.

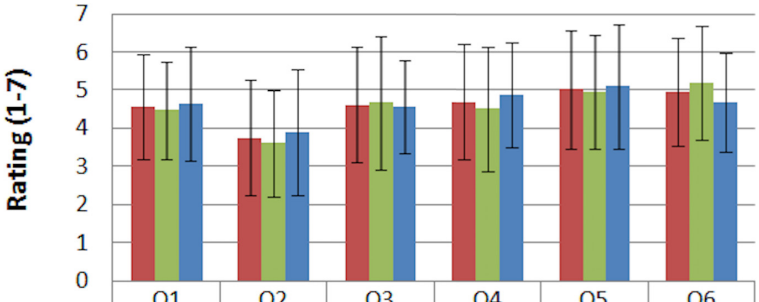
5.4. *Results*

5.4.1. *Questionnaires*

None of the participants had to be excluded due to their performance in the SAM test. Figure 7 presents the results from the questionnaires given to the participants. It shows the mean ratings of the participants' answers who heard the story told by Aesop using a TTS voice, using a human voice, as well as both. Between these two groups there are no significant differences regarding their ratings. However, in general, participants who listened to the story with the human voice rated the questions higher than the other participants. The questions are described in Sec. 4.1.2.

Table 5 presents the comparison of the three story appreciation factors — liking, engagement, empathizing the ratings of participants who heard the story told by a human voice or a TTS voice.

A one-way ANOVA revealed *significant differences* when comparing the scores of Q2 and Q5. These questions were related with the empathy generated between the participant and the storyteller ($F(39, 77) = 13.90$; $p < 0.001$). In the first question (Q2) the participants indicated if they felt sorry about the main character (beginning of the story), and the second question (Q5) if they felt happy (end of the story). The average of the ratings in these two questions indicates that at the end of the story the participants were more engaged with the story character than in the



	Q1	Q2	Q3	Q4	Q5	Q6
All Participants	4,55	3,75	4,63	4,70	5,03	4,95
TTS Voice	4,48	3,62	4,67	4,52	4,95	5,19
Human Voice	4,63	3,89	4,58	4,89	5,11	4,68

Fig. 7. Comparison of the mean ratings for each question in the questionnaire, where where vertical axis represents the participants' mean ratings based on a 7-point Likert scale; horizontal axis represents the mean ratings of each participant group on the Question 1 through Question 6.

Table 5. Comparison of the mean ratings between two groups in 7-point scale rating - M(SD).

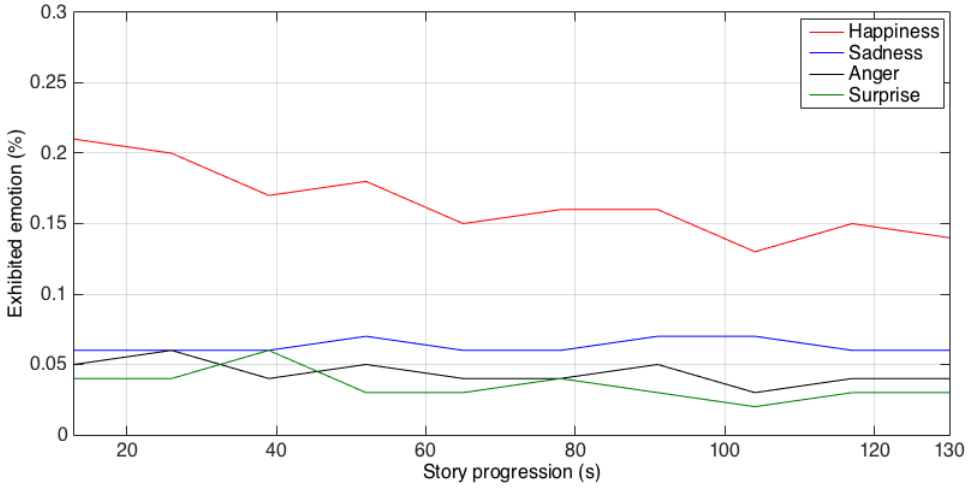
	All participants	TTS Voice	Human Voice
Liking (Q3+Q6)	4.79 (1.46)	4.93 (1.62)	4.63 (1.25)
Engagement (Q1+Q4)	4.63 (1.44)	4.5 (1.46)	4.76 (1.43)
Empathizing (Q2+Q5)	4.39 (1.53)	4.29 (1.45)	4.5 (1.65)

beginning. Since this questionnaire was done at the end of the experiment, when all the participants listened to the story both by Greta and by Aesop, through this questions alone, it is not possible to infer which one was a more effective storyteller.

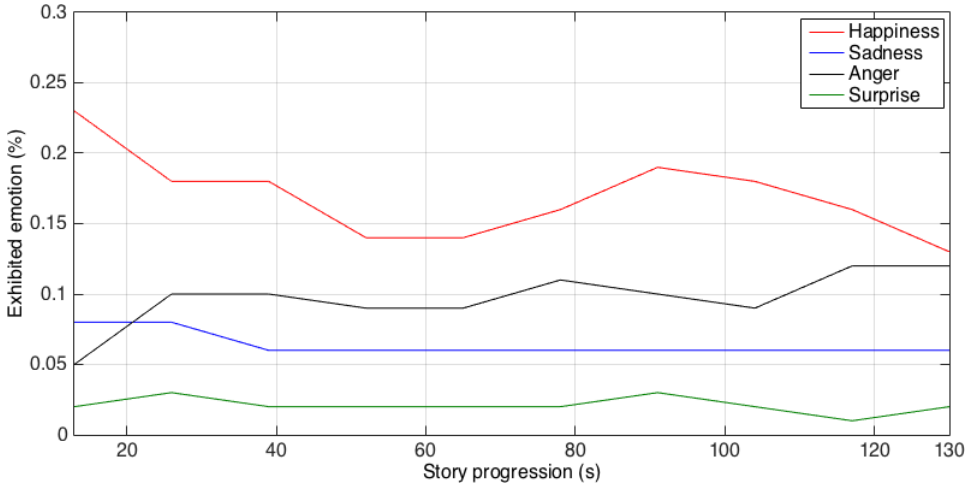
A one-way ANOVA test showed no significant differences when comparing Q1 with Q4 ($F(39, 77) = 0.643$; $p = 0.22$) and Q3 with Q6 ($F(39, 77) = 0.321$; $p = 0.99$). The pairs Q1/Q4, and Q3/Q6 were connected to the same appreciation factor, and thus it makes sense that no significant differences were found. Thus, our main finding here is that at the end of the story the participants reported feeling happy in a statistically significant stronger way as compared to their subjective reports of feeling sorry near the beginning of the story. This could be explained in multiple ways: either the magnitude of the sorry feeling was anyway smaller than the happiness at the end, or there is a memory effect which diminishes the remembered intensity of feeling sorry at the time of questionnaire answering, which takes place after the sorry feeling has been replaced by happiness. Also, and quite possibly, the empathy level of the listener towards the hero (the cracked pot) has increased through the progression of the story and the listener's familiarization with the hero, and this contributes positively to the intensification of the subjective reports of happiness, which is also the final emotional state of the story — in a sense, as the saying goes, all is well that ends well.

5.4.2. *Emotional analysis*

Figure 8 presents the emotional state shown by the participants during storytelling by Aesop either with a human and TTS voice. When observing this data, we see that both groups express a high percentage of happiness. The authors hypothesize that this might be related with the curiosity about the robot (subjective analysis from the experimenter during the trials). However, and regarding the collected data, on average no significant difference regarding the display of happiness was verified ($p > 0.05$) when comparing one group to the other. The comparison between the emotions displayed during storytelling with Aesop using a human voice and a TTS voice shows significant differences regarding anger ($p < 0.001$). Only in the first 20 seconds of the narrative there is no difference but with the story progress we see that the listeners of the story with a TTS voice became angrier. In average, the percentage of the exhibited emotion regarding sadness was higher and significantly different



(a)



(b)

Fig. 8. Percentage of the emotions shown in participants during storytelling by Aesop (Comparison between human and TTS voice). (a) Emotions shown in participants during storytelling with Aesop with a human voice and (b) Emotions shown in participants during storytelling with Aesop with a TTS voice.

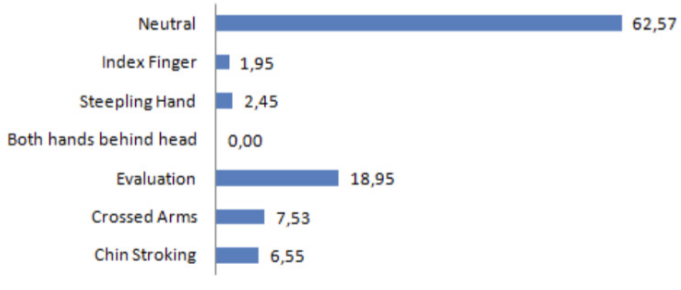
sadness ($p < 0.05$), in the story told by the robot using a human voice than a TTS voice. The same is verified for surprise ($p < 0.001$), thus illustrating the higher effectiveness of the human voice towards inducing emotions to the listener. *This addresses our research question RQ2: indeed, in this respect a human voice seems to be more effective than the current state-of-the art text-to-speech computer voice towards storytelling agents.*

5.4.3. *Blinking*

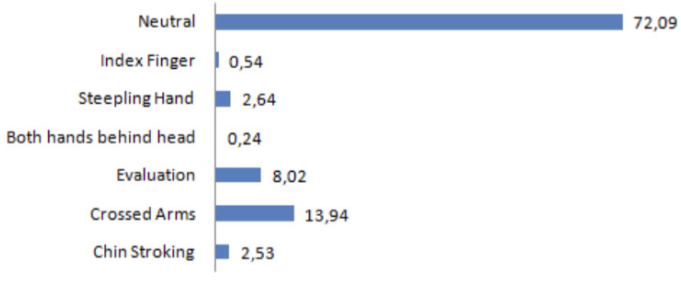
Average blink rate decreased along the story, as was the case in Study 1. We then compared the blink frequency while listening to the story told by different agents. The story was again divided into three parts, and a paired sample t-test was used for comparison. We found *significant differences* in the parts regarding the average of blinking rate ($p < 0.05$), when comparing the story being told by Greta and by Aesop. The average blinking rate is lower with Aesop than with Greta, indicating stronger listener engagement with the physical Robot as compared to the virtual Avatar, thus addressing our research question RQ1 - *indeed, in this respect a physical robot seems to be more effective than an avatar as an emotional storyteller.*

5.4.4. *Body language*

The videos from the experiments were used to evaluate the non-facial body language of the listener. Figures 9 and 10 represents the gestures performed using hands and arms. In this category, the predominant gestures were: crossed arms, evaluation, chin stroking, steepling hands, and neutral. Crossed arms indicate that the person disagrees or is not comfortable with the situation in front of him/her, and evaluation is self implied. Chin stroking indicates the listener is making a decision, while steepling hands is an illustration of confidence.[42] The ideal situation is that the

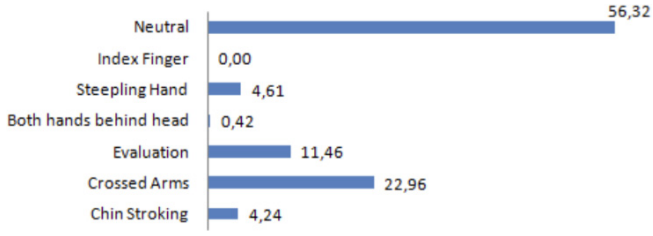


(a) Greta

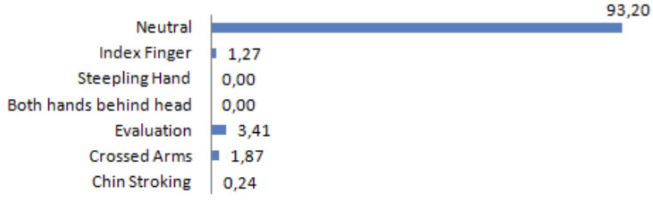


(b) Aesop

Fig. 9. Hand and arms gestures' percentages when the story was told by Greta and by Aesop.



(a) TTS



(b) Human voice

Fig. 10. Hand and arms gestures' percentages when the story was told by Aesop with a TTS and with a human voice.

participants' body language do not show behaviors indicating a negative attitude, as crossed arms, evaluation, chin stroking, and steepling hands.

ANOVA revealed *significant differences* between the time percentage the participants performed the aforementioned behaviors when the story was told either by the Greta avatar or by the Aesop robot, $F(6,7) = 50.92$; $p < 0.0001$. In fact, we verify that the participants have shown more neutral behaviors when listening to the story told by Aesop than when told by Greta. *Significant differences* were also found comparing between the behaviors, what was expected already observing the percentage of time dedicated to neutral behaviors. *Thus, in this respect too, regarding question RQ1, a physical robot seems to be preferable to a virtual agent as an emotional storyteller.*

When comparing the percentage of time the participants made these gestures when listening the story told by the robot either with a human voice or using a TTS voice, there are *significant differences* $F(6,7) = 9.18$; $p < 0.01$, and observing Fig. 10 we see that the story told using a human voice received more neutral behaviors. *Regarding question RQ2, for a storytelling agent human voice seems to be preferable to TTS.*

Gestures associated with legs were then examined. Crossed legs normally occurs with other negative gestures, and leg clamp is a sign of the tough-minded individual. Having the legs parallel to each other is an indication of a neutral opinion.[42] *Significant differences* were found when comparing the percentage of time the participants performed legs gestures when the story was told either by Greta or by Aesop, $F(2,3) = 47296.49$; $p < 0.001$ but no significant differences were found between the storytellers ($p > 0.005$). There were *significant differences* between the

percentage of time the participants made these gestures when listening the story told by the robot either with a human voice or using a TTS voice, $F(2,3) = 432.92$; $p = 0.002$. Thus, what we found is that there is no significant differences between the storytellers in this behavior. Regarding the leg movements of the participants, it is worth noting that given the positioning of the cameras and the table, in contrast to the movements of the rest of the body, the leg movements were neither always visible, nor were the legs absolutely free to move. However, we think it is worth reporting leg movements too, as the fact that some subjects chose to cross their legs even under these restrictive conditions might be a strong indication.

Regarding the head position category, we identified three different positions: neutral, interested and disapproval.[42] A two-way ANOVA showed *significant differences* between the percentage of time the participants showed the previously mentioned behaviors when the story was told either by Greta or by the robot, $F(2,3) = 49.48$; $p = 0.003$, but no significant difference was found between the storytellers in this behavior. The same is verified when comparing the percentage of time the participants made these gestures when listening the story told by the robot either with a human voice or using a TTS voice, $F(2,3) = 78,71$; $p = 0.01$. Thus, we found no significant differences between the storytellers in this behavior.

Thus, from the participants' non-verbal response analysis, we would like to highlight that the fact that the participants were seated in front of the storyteller having a table between them, did not allow the participants to freely exhibit behaviors from their legs. In addition, focus should be given to hands and arms gestures, where significant differences were found both between groups and between behaviors.

To recap: significant differences across groups (robot versus avatar and human voice versus TTS) were found in the gestures involving head and arms, namely: crossed arms, evaluation, chin stroking, steepling hands, and neutral. These provided yet further results towards answering our questions RQ1 and RQ2, and supported the higher effectiveness of robot embodiment as compared to virtual avatars for storytelling, and the higher effectiveness of the human voice as compared to a synthesized voice for such storytelling.

6. Discussion and Future Work

Towards our ultimate goal of emotional storytelling agents that can observe their listener and adapt their style in order to maximize their effectiveness, we have set out three main research questions, which we investigated through experiments utilizing questionnaires, automated analysis of facial expressions and blink rate, and manual analysis of non-verbal gestures.

The basic concern of Research Question 1 was *embodiment*: is a physical robot preferable to a virtual avatar for emotional storytelling? Our results indeed indicate that it is. The average blinking rate is lower with Aesop than with Greta, indicating stronger listener engagement with the physical Robot as compared to the virtual Avatar. Non-verbal gestures involving head and arms (crossed arms, evaluation, chin

stroking, steepling hands, and neutral) also support this claim, as for example the participants have shown more neutral behaviors when listening to the story told by Aesop than when told by Greta.

As a side question, RQ1a, we asked whether there are differences in narrative engagement across different parts of the story. As reported, blink rate data as well as questionnaire answers indicated that the apparent engagement was higher towards the climactic end of the story.

Research Question 2 was concerned with the choice of storytelling *voice*: is the state-of-the art of text-to-speech synthesizers good enough, or is a human voice still preferable? Our results indicate that a human voice is still preferable: both in terms of exhibited listener facial expressions indicating emotions during storytelling, as well as in terms of non-verbal gestures involving head and arms, we reached statistically significant difference twice. Still, text to speech technology has not reached the right level of maturity to compete with human voices for storytelling agents.

Finally, we reach the most interesting Research Question 3. Will empathy be exhibited by the listener's emotional reactions to the story? And at an even wider scope, a very important yet interesting question arises: what is the relation between the postulated emotional trajectories of different story characters, the emotional trajectory of the storyteller, and that of the listener? Is it a simple equality relation or a fixed one-to-one mapping?

Let us start with empathy. Our results of Figs. 4 and 5 indicate that indeed there is empathy between the storyteller and the listener, as evidenced by the equality of valence between the narrator's facial expressions (corresponding to the dominant emotions of the story derived as described) and the listener's facial expressions, over time. Even more so, by looking at Figs. 2 and 3 and the commentary that follows them, it becomes evident that an interesting yet complex relation of emotional "harmonies": In analogy to musical harmony, where chords are formed through the vertical co-temporality of notes played by various voices, while the melodic lines of these voices are unfolding in parallel over time, one can postulate emotional "harmonies" among the participants in narrative. That is, as the emotional lines of different narrative participants (story characters, storyteller, listener) evolve over time, one can postulate the vertical "chords" that are formed by them.

One could envision that there exist many interesting patterns, constraints, and relations, both on the temporal dynamics of the emotional lines of each voice (character, storyteller, listener) in narrative, as well as on the vertical relations of the co-temporal content across the lines. For example, some such indications that we have observed in Figs. 2 and 3 indicate that there exist cases where there is equality of voices, such as event E3(happy) = E3'(happy), and other cases where there is no simple fixed one-on-one relation E1(sadness) - E1'(anger), while E2(sadness) - E2'(anger and sadness). Of course, one can postulate explanation behind these patterns in terms of both the temporal progression of the story (horizontal aspect) as well as the co-temporal reactions to emotional states. This is a plausible account of the observed emotional melodies and harmonies of the storyteller and the listener as

just the beginnings of a potential theory connecting concepts of music (melody and harmony, and possibly dissonance) with emotional trajectories of multiple narrative participants evolving over time, which opens up exciting avenues for further research along with a group of studies to explore the meaning and narrativity in music.[51,52]

Another interesting comment is the following. The results from the questionnaires in Study 2 (Sec. 5.4.1) suggest the participants were empathizing less with the main character in the beginning of the story (Q2), when it was feeling bad about itself. However, Q5 scores show us that in the end of the story the participants shared its joy more strongly, and as we discussed above, one possible explanation is that an empathetic connection between the listener and the hero was being formed and strengthened along the story. This is a good indication that the way the story was encoded in the Study 1 promoted empathy. In the first part of the story, the listener might had been building a mental model of the story and its entities, and establishing a connection with the storyteller and the characters. While the situation model of the story was being created in the listener's mind,[53,54] the mental models of the characters represented in the listener's mind started as generic models of human characters, without specific information attached to them. However, the more the listener learned about them as the story narration progressed, they started becoming more specific. For stronger empathizing, the story needed to have progressed enough in order to know enough about the character in order to empathize strongly with him or her, and with his traits, and history. This result is congruent with the findings of RQ3, which indicate empathy at the valence level between storyteller and listener, again supporting the effectiveness of our storytelling agent.

Numerous future extensions exist: First, in the current study, a single story was only used. A small yet varied repertoire of stories of similar duration would be beneficial, and the variation could help obtain more cases of interesting emotional line progressions and harmonies. Furthermore, if one extends to longer stories, then other phenomena, such as total loss of narrative attention, might become apparent. Also, we could start modulating not only the facial expressions of the storytelling agent, but also his voice prosody, temporal parameters such as pauses, etc. Furthermore, the level of the anthropomorphicity of the storyteller could be varied, and also the age groups of listeners. Most importantly, the above notes on the emotional "harmonies" and "melodies" that take place during storytelling across characters, storyteller, and listener, could be turned into a formal theory, and targeted experiments performed to further inform it.

7. Conclusion

Towards our ultimate goal of real-time adaptive emotional storytelling agents, we have presented experiments and results towards answering three important research questions, termed RQ1-RQ3, and we have furthermore provided a discussion including the early steps of a theory of emotional "melodies" and "harmonies" in analogy to their musical counterparts.

We started by noting that numerous aspects of narration, including facial expressions and prosody, are important towards creating interesting and effective artificial storytellers, either robotic or virtual. Then, we noted that ideally, the manner of narration should be adaptable in real-time during storytelling, based on feedback derived from non-verbal cues arising from the listener, given the differences in engagement and preferences of different listeners at different times. But in order to start experimenting with either off-line or on-line real-time adaptation, a number of fundamental research questions came up, whose answers are highly important: First, is a physically embodied agent preferable to a virtual agent or a voice-only narration? Second, does a human voice have an advantage over a synthesized voice? Third, how should the emotional trajectory of the characters in a story be related to a storyteller's facial expressions during storytelling time, and how does this correlate with the emotions on the faces of the listeners?

In this paper, we provided empirical answers to the above questions, through two specially designed studies, during which we observed the reactions of experimental subjects to artificial storytellers, through a combination of instruments: special questionnaires, manual body language annotation, and automated facial expression and blink analysis. The results indicate that the physically embodied robot produces more narrative attention to the listener as compared to a virtual embodiment, that a human voice is preferable over the current state of the art of text-to-speech, and that the empathizing of the listener is evident through its facial expressions. Most importantly, it became apparent that there is a complex yet interesting relation between the emotion lines of the story, the facial expressions of the narrating agent, and the emotions of the listener, and the beginnings of a theory of emotional "melodies" across time and "harmonies" agents were introduced in our discussion, where further future steps were also discussed. This work constitutes an important step towards emotional storytelling robots that can observe their listener and adapt their style in order to maximize their effectiveness, thus enabling the further beneficial entry of robots towards enhancing our everyday life.

Acknowledgments

To the IRSS2013 participants, to the Portuguese Foundation (FCT) for funding through the R&D project RIPD/ADA/109407/2009, SFRH/BD/71600/2010, and FCOMP-01-0124-FEDER-022674. This work was also supported in part by the Hongik University new faculty research support fund.

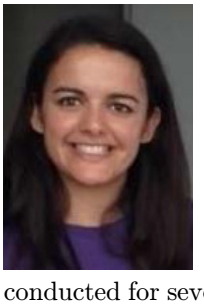
References

1. R. C. Roney, Back to the basics with storytelling, *Reading Teacher* **42**(7) (1989) 520–523.
2. G.-D. Chen *et al.*, A survey on storytelling with robots, in *Edutainment Technologies. Educational Games and Virtual Reality/Augmented Reality Applications* (Springer, 2011), pp. 450–456.

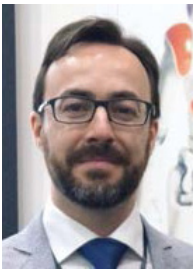
3. C.-H. J. Lee, C.-Y. I. Jang, T.-H. D. Chen, J. Wetzel, Y.-T. B. Shen and T. Selker, Attention meter: A vision-based input toolkit for interaction designers, in *CHI'06 Extended Abstracts on Human Factors in Computing Systems* (ACM, 2006), pp. 1007–1012.
4. K. Drummond and R. Hopper, Back channels revisited: Acknowledgment tokens and speakership incipiency, *Res. Language Soc. Interact.* **26**(2) (1993) 157–177.
5. D. Tang, B. Yusuf, J. Botzheim, N. Kubota and C. S. Chan, A novel multimodal communication framework using robot partner for aging population, *Expert Syst. Appl.* **42**(9) (2015) 4540–4555, http://dx.doi.org/10.1016/j.eswa.2015.01.016.
6. M. Mancas, V. P. Ferrera, N. Riche and J. G. Taylor, *From Human Attention to Computational Attention: A Multidisciplinary Approach*, 1st edn. (Springer Publishing Company, Incorporated, 2016).
7. G. A. Miller, The magical number seven plus or minus two: Some limits on our capacity for processing information, *Psychol. Rev.* **63**(2) (1956) 81–97.
8. M. Csikszentmihalyi, *Flow: The Psychology of Optimal Experience*, 1st edn. (Harper and Row, New York, 1990).
9. R. Busselle and H. Bilandzic, Measuring narrative engagement, *Media Psychol.* **12**(4) (2009) 321–347.
10. K. Oatley, A taxonomy of the emotions of literary response and a theory of identification in fictional narrative, *Poetics* **23**(1) (1995) 53–74.
11. R. W. Levenson and A. M. Ruef, Empathy: A physiological substrate, *J. Person. Soc. Psychol.* **63**(2) (1992) 234.
12. J. Decety and P. L. Jackson, The functional architecture of human empathy, *Behav. Cogn. Neurosci. Rev.* **3**(2) (2004) 71–100.
13. J.-J. Cabibihan, W. C. So and S. Pramanik, *Human-Recognizable Robotic Gestures*, 2012.
14. A. R. Bentivoglio, S. B. Bressman, E. Cassetta, D. Carretta, P. Tonali and A. Albanese, Analysis of blink rate patterns in normal subjects, *Movement Disorders* **12**(6) (1997) 1028–1034.
15. S. Shultz, A. Klin and W. Jones, Inhibition of eye blinking reveals subjective perceptions of stimulus salience, *Proc. Nat. Acad. Sci.* **108**(52) (2011) 21270–21275.
16. A. Silva, M. Vala and A. Paiva, Papous: The virtual storyteller, in *Intelligent Virtual Agents* (Springer, 2001), pp. 171–180.
17. A. Silva, G. Raimundo and A. Paiva, Tell me that bit again... bringing interactivity to a virtual storyteller, in *Virtual Storytelling. Using Virtual RealityTechnologies for Storytelling* (Springer, 2003), pp. 146–154.
18. F. Charles, S. Lemercier, T. Vogt, N. Bee, M. Mancini, J. Urbain, M. Price, E. André, C. Pélachaud and M. Cavazza, Affective interactive narrative in the callas project, in *Virtual Storytelling. Using Virtual Reality Technologies for Storytelling* (Springer, 2007), pp. 210–213.
19. P. Bleackley, S. Hyniewskab, R. Niewiadomski, C. Pelachaud and M. Price, Emotional interactive storyteller system, *Proc. Int. Conf. Kansei Engineering and Emotion Research*, Paris, France, 2010.
20. C. Plaisant, A. Druin, C. Lathan, K. Dakhane, K. Edwards, J. M. Vice and J. Montemayor, A storytelling robot for pediatric rehabilitation, in *Proc. Fourth Int. ACM Conf. Assistive Technologies* (ACM, 2000), pp. 50–55.
21. C. J. Wong, Y. L. Tay, R. Wang and Y. Wu, Human-robot partnership: A study on collaborative storytelling, in *The Eleventh ACM/IEEE Int. Conf. Human Robot Interation, HRI 2016*, Christchurch, New Zealand, March 7–10, 2016, pp. 535–536, http://dx.doi.org/10.1109/HRI.2016.7451843.
22. K. T. Ying, S. B. M. Sah and M. H. L. Abdullah, Personalised avatar on social stories and digital storytelling: Fostering positive behavioural skills for children with autism

spectrum disorder, in *2016 4th Int. Conf. User Science and Engineering (i-USEr)*, August 2016, pp. 253–258.
23. B. Mutlu, J. Forlizzi and J. Hodgins, A storytelling robot: Modeling and evaluation of human-like gaze behavior, in *6th IEEE-RAS Int. Conf. Humanoid Robots* (IEEE, 2006), pp. 518–523.
24. M. Sugimoto, A mobile mixed-reality environment for children's storytelling using a handheld projector and a robot, *IEEE Trans. Learning Technologies* **4**(3) (2011) 249–260.
25. K. Ryokai, M. J. Lee and J. M. Breitbart, Children's storytelling and programming with robotic characters, in *Proc. Seventh ACM Conf. Creativity and Cognition*, ACM, 2009, pp. 19–28.
26. C. R. Ribeiro, C. P. Coutinho and M. F. Costa, Robotics in Child Storytelling, 2009.
27. T. Munekata, Y. Fujita and T. Nishizawa, New learning environment for enhancing storytelling activities of children with intellectual disabilities/autism using a personal robot in the classroom, in *Semantic Scholar*, 2009.
28. J. Kennedy, P. Baxter and T. Belpaeme, Comparing robot embodiments in a guided discovery learning interaction with children, *Int. J. Soc. Robot.* **7**(2) (2015) 293–308, http://dx.doi.org/10.1007/s12369-014-0277-4.
29. I. Poggi, C. Pelachaud, F. de Rosis, V. Carofiglio and B. De Carolis, Greta. a believable embodied conversational agent, in *Multimodal Intelligent Information Presentation* (Springer, 2005), pp. 3–25.
30. N. Mavridis and D. Hanson, The ibnsina center: An augmented reality theater with intelligent robotic and virtual characters, in *The 18th IEEE Int. Symp. Robot and Human Interactive Communication, 2009. RO-MAN 2009* (IEEE, 2009), pp. 681–686.
31. L. Riek, N. Mavridis, S. Antali, N. Darmaki, Z. Ahmed, M. Neyadi and A. Ketheri, Ibn sina steps out: Exploring arabic attitudes toward humanoid robots, *Proc. 2nd Int. Symp. New Frontiers in Human–Robot Interaction*, AISB, Leicester, Vol. 1, 2010.
32. N. Mavridis, M.-S. Katsaiti, S. Naef, A. Falasi, A. Nuaimi, H. Araifi and A. Kitbi, Opinions and attitudes toward humanoid robots in the middle east, *AI & Soc.* **27**(4) (2012) 517–534.
33. M. Mancini and C. Pelachaud, The fml-apml language, in *Proc. Workshop on FML at AAMAS*, Vol. 8, 2008.
34. SHORE, Unknown title, http://www.iis.fraunhofer.de/en/bf/bsy/produkte/shore/.
35. FaceAPI, http://www.seeingmachines.com/product/faceapi/.
36. M. M. Bradley and P. J. Lang, Measuring emotion: The self-assessment manikin and the semantic differential, *J. Behav. Therapy Exp. Psychiatry* **25**(1) (1994) 49–59.
37. M. Pantic and L. J. M. Rothkrantz, Automatic analysis of facial expressions: The state of the art, *IEEE Trans. Pattern Analysis and Machine Intelligence* **22**(12) (2000) 1424–1445.
38. K. Scherer, Psychological models of emotion, in *The Neuropsychology of Emotion*, J. C. Borod (ed.) (Oxford University Press, New York, 2000), pp. 137–162.
39. J. Russell, A circumplex model of affect, *J. Personality Soc. Psychol.* **39**(6) (1980) 1161–1178.
40. L. Noldus, The observer: A software system for collection and analysis of observational data, *Behav. Res. Methods, Instrum. Comput.* **23**(3) (1991) 415–429.
41. C. Christoforou, N. Mavridis, E. L. Machado and G. Spanoudis, Android tele-operation through brain-computer interfacing: A real-world demo with non-expert users, in *Proc. Int. Symp. Robotics and Intelligent Sensors (IRIS)*, 2010.
42. A. Pease, Body language, 2012.

43. P. Ekman and W. Friesen, Head and body cues in the judgement of emotion: A reformulation, *Percept Motor Skill* **24** (1967) 711–724.
44. C. Busso, Z. Deng, M. Grimm, U. Neumann and S. Narayanan, Rigid head motion in expressive speech animation: Analysis and synthesis, *IEEE Audio Speech Language Process* **15** (2007) 1075–1086.
45. H. Wallbott, Bodily expressions of emotion, *Eur. J. Soc. Psychol.* **28** (1998) 879–896.
46. J. Fast, Body language, *Pocket, 1998*.
47. L. Carmichael, S. Roberts and N. Wessell, A study of the judgment of manual expression as presented in still and motion pictures, *J. Soc. Psychol.* **8** (1937) 115–142.
48. J. Reilly, M. L. McIntire and H. Seago, Affective prosody in american sign language, *Sign Language Studies* **75** (1992) 113–128.
49. A. Samadani, B. DeHart, K. Robinson, D. Kulic, E. Kubica and R. Gorbet, A study of human performance in recognizing expressive hand movements, *IEEE Int. Symp. RO-MAN*, 2011, pp. 93–100.
50. R. Bakeman and J. Gottman, *Observing Interaction: An Introduction to Sequential Analysis* (Cambridge University Press, 1997).
51. L. B. Meyer, *Emotion and Meaning in Music* (The University of Chicago Press, Chicago, 1956).
52. F. E. Maus, Narrative, drama, and emotion in instrumental music, *J. Aesthetics Art Criticism* **55**(3) (1997) 293–303.
53. R. A. Zwaan and G. A. Radvansky, Situation models in language comprehension and memory, *Psychol. Bull.* **123**(2) (1998) 162.
54. N. Mavridis, Grounded situation models for situated conversational assistants, Ph.D. dissertation, MIT, 2007, available online at DSpace.



Sandra Costa studied Electronical Engineering at the University of Minho, Portugal where she obtained her master degree investigating the use of robotics in low-functioning young adults with Autism Spectrum Disorders (ASD) Therapy. When she finishes her master, she chose to continue in the same field of research for her doctorate. In 2015, she was awarded her Ph.D. for the work developed regarding affective robotics for socio-emotional skills development in children with ASD, where a study was conducted for several months with three groups of 15 high-functioning children with ASD, with significant results for the use of a humanoid robot as a useful tool to develop socio-emotional skills, due to the engagement and positive learning outcome observed. Presently, Sandra works for Bosch Car Multimedia, in Portugal in the Autonomous Driving Sensing business.



Alberto Brunete received his M.S. degree in Telecommunication Engineering and his Ph.D. degree in Robotics and Automation from the Technical University of Madrid (UPM) (Spain) in 2000 and 2010, respectively. He has been Senior Researcher and Technical Manager at the Research Center for Smart Buildings and Energy Efficiency (CeDInt-UPM) and Visiting Professor in the Department of System Engineering and Automation at the Carlos III University (Spain). He is currently Assistant Professor at the Technical University of Madrid and researcher at the Centre for Automation and Robotics (CAR UPM-CSIC). His main research activities are related to robotics and smart environments (ambient intelligence and ambient assisted living). In 2016 he has won the Spanish prize "ABC Solidario".

Byung-Chull Bae received the B.S. degree in 1993 and the M.S. degree in 1998 from Korea University, South Korea, and the Ph.D. degree from North Carolina State University, Raleigh, NC, USA, in 2009. He is currently an Assistant Professor at School of Games, Hongik University, Sejong, South Korea. He has worked at LG Electronics and Samsung Electronics as a research engineer, and worked for IT University of Copenhagen, Denmark, as a visiting scholar and a part-time lecturer. His research interests include interactive storytelling, affective computing, and game AI.

Nikolaos Mavridis received his Ph.D. from the Massachusetts Institute of Technology, after receiving his M.Sc. from UCLA and M.Eng. from the Aristotle University of Thessaloniki. He has served as faculty at numerous universities, including New York University and UAEU, and as a researcher at NCSR Demokritos. He has also served in numerous service roles, such as the vice-chair of the Hellenic Artificial Intelligence Society, the founding chair of IEEE UAE Robotics and Automation Society, as a judge for the Dubai Prime Minister's Office "Robotics and AI for good" competition, and more. His research interests include Human–Robot Interaction, Artificial Intelligence, and Cognitive Systems.