International Journal of Social Robotics (2019) 11:555–573
https://doi.org/10.1007/s12369-019-00524-z

Multimodal Integration of Emotional Signals from Voice, Body, and Context: Effects of (In)Congruence on Emotion Recognition and Attitudes Towards Robots

Christiana Tsiourti[1] · Astrid Weiss[2] · Katarzyna Wac[3] · Markus Vincze[1]

Accepted: 18 January 2019 / Published online: 4 February 2019

Abstract

Humanoid social robots have an increasingly prominent place in today's world. Their acceptance in social and emotional human–robot interaction (HRI) scenarios depends on their ability to convey well recognized and believable emotional expressions to their human users. In this article, we incorporate recent findings from psychology, neuroscience, human–computer interaction, and HRI, to examine how people recognize and respond to emotions displayed by the body and voice of humanoid robots, with a particular emphasis on the effects of incongruence. In a social HRI laboratory experiment, we investigated contextual incongruence (i.e., the conflict situation where a robot's reaction is incongrous with the socio-emotional context of the interaction) and cross-modal incongruence (i.e., the conflict situation where an observer receives incongruous emotional information across the auditory (vocal prosody) and visual (whole-body expressions) modalities). Results showed that both contextual incongruence and cross-modal incongruence confused observers and decreased the likelihood that they accurately recognized the emotional expressions of the robot. This, in turn, gives the impression that the robot is unintelligent or unable to express "empathic" behaviour and leads to profoundly harmful effects on likability and believability. Our findings reinforce the need of proper design of emotional expressions for robots that use several channels to communicate their emotional states in a clear and effective way. We offer recommendations regarding design choices and discuss future research areas in the direction of multimodal HRI.

Keywords Social robots · Human–robot interaction · Robot emotions · Multi-modal interaction · Body language · Believability

1 Introduction

In the last years, growing interest has been observed in the development of socially intelligent robots, which are envisioned to interact with humans in a variety of social and emotional roles, such as household assistants, companions for children and elderly, partners in industries, guides in public spaces, educational tutors at school and so on [1]. There is accumulating evidence that expressive robots, equipped with the ability to show human-like emotions, are rated as more likable and humanlike, and lead to higher engagement and more pleasurable interactions [2–5]. Additionally, trust, acceptance, and cooperation with a robot are dependent on the match between the social context of the situation and the emotional behaviour of the robot [6,7]. Therefore, understanding how people perceive and interact with emotional robots is crucial, given the growing deployment of these robots in social settings.

✉ Christiana Tsiourti
christiana.tsiourti@tuwien.ac.at

Astrid Weiss
astrid.weiss@tuwien.ac.at

Katarzyna Wac
wac@di.ku.dk

Markus Vincze
vincze@acin.tuwien.ac.at

[1] Automation and Control Institute (ACIN), Vision4Robotics Group,TU Wien, Guhausstrae 27, 1040 Vienna, Austria

[2] Institute of Visual Computing and Human-Centered Technology, Human–Computer Interaction (HCI) Group, TU Wien, Argentinierstrae 8/E193-5, 1040 Vienna, Austria

[3] Human-Centered Computing Section, Quality of Life Technologies Group, University of Copenhagen, Emil Holms Kanal 6, 2300 Copenhagen, Denmark

Humans are experts in social interaction. During face-to-face social interactions, the human sensory system uses multimodal analysis of multiple communication channels to recognize another party's affective and emotional states [8]. A channel is a communication medium; for example, the auditory channel carries speech and vocal intonation communicative signals, and the visual channel carries facial expressions and body language signals. A modality is a sense, used to perceive signals from the outside world (i.e., sight, hearing) [9]. Engaging in a "routine" conversation is a rather complex multimodal task; a human must carefully attend to and decipher cues encountered in different sensory modalities and several communication channels at once. Multimodality ensures that the analysis of affective information is highly flexible and robust. Failure of one channel is recovered by another channel and information in one channel can be explained by information in another channel (e.g., a facial expression that might be interpreted as a smile will be interpreted as a display of sadness if at the same time we see tears and hear weeping) [8].

To be effective social interaction partners, robots must also exploit several channels (i.e., auditory, visual) and mechanisms (e.g., body posture, facial expressions, vocal prosody, touch, gaze) to communicate their internal emotional states and intentions in an authentic and clear way [10,11]. Many researchers, explore the design space of anthropomorphic or zoomorphic robots equipped with expressive faces (e.g., [5,12–16]), emotional voices (see [17] for a survey), body language (e.g., [18–21]), and other features and capacities to make human–robot social interactions more human-like. While initially, Human–Robot Interaction (HRI) research on the emotional expressions of robots had largely focused on single modalities in isolation [22], more recently, researchers have begun to integrate multiple channels, in order to approach the richness of human emotional communication. For instance, HRI studies have examined the perception of emotional expressions involving faces [23], faces and gestures [24], gestures and voices [25] and face-voice-gesture [23] combinations.

The results of these studies show that recognition accuracy as well as attitudes towards robots, such as expressiveness [23], likability [25] and trust [6], increase when multiple channels are used to convey congruent emotional information. Although it may not be possible to incorporate all features of human communication into robots (due to the complexity of the phenomenon), the affect-expression capability of humans can serve as the "gold standard" and a guide for defining design recommendations for multimodal expression of human-like affective states. The importance of congruence in multimodal emotional expressions of robots remains largely under-explored. Two of the primary channels which robots use to express emotion multimodally are the auditory and the visual channel. However, the importance of using an appropriate combination of audio-visual stimuli when conveying emotions remains under-explored in HRI. Both the psychological and Human–Computer Interaction (HCI) literature suggest that people favour congruence (also known as consistency) over incongruence (also known as inconsistency). For instance, studies with Embodied Conversational Agents (ECAs) [26–30] that have presented congruent and incongruent auditory and visual stimuli at the same time, showed that emotional information conveyed in one modality (i.e., vocal prosody, facial expressions) influences the processing of emotional information in the other modality, and congruent emotional information across auditory and visual channels tends to facilitate emotion recognition. Conversely, incongruent emotional responses can result in adverse consequences on user ratings (i.e., trust, likability, expressiveness) towards ECAs [30]. Does the same apply during interactions with robots? For example, what do people perceive if they observe a robot with a happy body posture combined with a concerned voice? Do people base their perceptions of the emotion on one channel more than another? That is, does either the visual or audio channel dominate in perceptions of the emotional expression, or are they both essential? Additionally, what is the impact of incongruent emotional expressions on people's attitudes towards robots?

In this article, we aim at investigating the multimodal perception of emotions of humanoid robots, in the context of social interactions with humans. We draw insights and perspectives about the multisensory integration of congruent and incongruent emotional information from the fields of HRI, HCI, psychology, and neuroscience. We investigate how people recognize emotions expressed by a humanoid robot multimodally, via two different modalities: the body (i.e., head, arms and torso position and movement) and the voice (i.e., pitch, timing, loudness and non-verbal utterances). We consider two distinct cases of incongruence, namely, contextual incongruence and cross-modal incongruence. The first case refers to the conflict situation where the robot's reaction is incongrous with the socio-emotional context of the interaction (e.g., a robot expresses happiness in repsonse to a sad situation). The second case refers to the conflict situation where an observer receives incongruous emotional information across the auditory (robot's vocal prosody) and visual (robot's whole-body expressions) modalities (e.g., a robot expresses sad voice and happy body postures). We investigate the effects of contextual incongruence and cross-modal incongruence on people's ability to recognize the emotional expressions of a robot, as well as on people's attitudes towards a robot (i.e., believability, perceived intelligence and likability). Specifically, we address the following research questions:

1. How are voice (i.e., pitch, timing, loudness, and non-verbal utterances) and body (i.e., head, arms and torso position and movement) expressions of a humanoid robot perceived when presented simultaneously with congruent and incongruous socio-emotional context?
2. How are voice and body expressions of a humanoid robot perceived when presented simultaneously in congruent and incongruent multimodal combinations?
3. What impact does incongruence have on people's perceptions of the robot, in terms of believability, perceived intelligence, and likability?

The rest of this article is organized as follows: we start by discussing the importance of multisensory interaction in human social interactions, and highlight research which has examined multisensory integration effects using behavioural experiments, functional neuroimaging (fMRI) and electroencephalography (EEG) measurements in psychology. We then discuss the importance of congruence in the context of HRI and detail relevant research that has already investigated this space. Following this, we describe a social HRI laboratory experiment which we conducted to investigate our research questions. A discussion of the results is then provided, followed by a set of guiding principles for future research towards the design of multimodal emotional expressions for humanoid robots.

2 Background and Related Work

2.1 Multisensory Interaction (MI) Research in Psychology and Neuroscience

MI refers to the processes by which information arriving from one sensory modality interacts with, and sometimes biases, the perception of cues presented in another modality, including how these sensory inputs are combined to yield a unified percept [31–33]. MI effects have been studied using behavioural experiments, functional neuroimaging (fMRI) and electroencephalography (EEG) measurements with faces and voices [34–36], faces and bodies [37,38], body expression and voices [38,39], and body and sound stimuli [40]. The results suggest strong bidirectional links between emotion detection processes in vision and audition. Additionally, there is accumulating evidence that integration of different modalities, when they are congruent and synchronous, leads to a significant increase in emotion recognition accuracy [41]. However, when information is incongruent across different sensory modalities, integration may lead to a biased percept, and emotion recognition accuracy is impaired [41].

Perception of Emotion From Face and Voice Previous MI research has mainly investigated the perception of emotional face-voice combinations [34–36]. For example, de Gelder and Vroomen [34] presented participants with static images of facial expressions that were morphed on a continuum between happy and sad, combined with a short spoken sentence. This sentence had a neutral meaning but was spoken in either a happy or sad emotional tone of voice. Participants were instructed to attend to and categorize the face, and to ignore the voice, in a two-alternative forced-choice task. The results showed a clear influence of the task-irrelevant auditory modality on the target visual modality. When asked to identify the facial expression, while ignoring the simultaneous voice, participants' judgments were nevertheless influenced by the tone of the voice and vice versa.

Perception of Emotion From Body and Voice More recently, researchers examining the integration of emotional signals from different modalities have started to pay attention to bodily expressions. A handful of studies have examined body expression and face combinations [37,38], body and sound stimuli (e.g., [40]), as well as body and voice combinations (e.g., [38,39]). The results follow a similar pattern to studies of emotional faces and voices. For example in [38], the authors used a similar paradigm as de Gelder and Vroomen [34] but tested for the effect of body expressions. Participants were presented with static images of whole-body expressions combined with short vocal verbalizations. The results indicate that the perceived whole-body expression influenced the recognition of vocal prosody. When observers make judgments about the emotion conveyed in the voice, recognition was biased toward the simultaneously perceived body expression.

Perception of Emotion From Face and Contextual Information A few studies have studied interactions between emotional faces paired with contextual information (e.g., [42–44]). Such experimental paradigms reveal that emotion perception is not driven by information in the face, body or voice alone but also derives from contextual information in the environment, such as the emotion-eliciting situation where the perceived emotion occurs, or the observer's current emotional state (e.g., [43,45]). Using fMRI, Mobbs et al. [42] demonstrated that pairing identical faces with either neutral or emotionally salient contextual movies results in both altered attributions of facial expression and mental-state. In this study, evaluators were presented with 4 s of a movie (positive, negative, and neutral) and were then shown an image of an emotional face (happy, fear, and neutral). Evaluators rated the combined presentations. Faces presented with a positive or negative context were rated significantly differently than faces presented in a neutral context. Furthermore, fMRI data showed that pairings between faces and emotional movies resulted in enhanced BOLD responses in several brain regions which may act to guide appropriate choices across altering contexts. In another study [44], situational cues in the form of short vignettes were found to influence the labelling of subsequently presented facial expressions (e.g., a sad face was labelled as “sad” when presented in isolation but was labelled as “disgust” when preceded by a disgust-related vignette). Niedenthal et al. [43] investigated congruent or incongruent facial expressions and surrounding context pairs. When the surrounding emotional context did not match the facial expression, observers often reinterpreted either the facial expression (i.e., the face does not reveal the person’s real feelings) or altered their interpretation of the contextual situation.

2.2 Multisensory Interaction Research in HCI

Research on the integration of multiple emotional signals is a relatively new topic in the areas of HCI and HRI. However, accumulating evidence from recent studies with anthropomorphic ECAs (e.g., [26–29]) and ro-bots (e.g., [6,23–25, 46,47]) shows that MI effects are also highly pronounced in the perception of synthetic emotional expressions. A number of HCI studies have used ECAs to investigate the experimental conflict situation where an observer receives incongruent information from two different sensory modalities (i.e., vocal prosody, facial expressions or body expressions). For example, Clavel et al. [26] studied the role of face and body in the recognition of emotional expressions of an ECA, using congruent emotional expressions, where the emotions expressed by the ECA’s face and body matched, and incongruent expressions, where the emotions expressed across the two modalities were mismatched. Their results showed that emotion recognition improves when the facial expression and body posture are congruent. The authors also reported that emotional judgments were primarily based on the information displayed by the face, although recognition accuracy improved when congruent postures were presented. Mower et al. [28] investigated the interaction between the face and vocal expressions and their role in the recognition of an ECA’s emotional expressions. In this study, the authors combined human emotional voices with synthetic facial expressions of an ECA, to create ambiguous and conflicting audio-visual pairs.

The results indicated that observers integrate natural audio cues and synthetic video cues only when the emotional information is congruent across the two channels. Due to the unequal level of expressivity, the audio was shown to bias the perception of the evaluators. However, even in the presence of a strong audio bias, the video data were shown to affect human perception. Taken together, the abovementioned results from ECA studies that have presented congruent and incongruent auditory and visual stimuli at the same time, suggest that emotional information conveyed in one modality influences the processing of emotional information in the other modality and that congruent emotional information tends to facilitate emotion recognition.

Studies with ECAs also report that congruence is associated with increased expressiveness (e.g., [23]), likability (e.g., [25,27]) and trust towards the agents (e.g., [6,23],). Creed et al. [27] investigated the psychological impact of a virtual agent’s mismatched face and vocal expressions (e.g., a happy face with a concerned voice). The mismatched expressions were perceived as more engaging, warm, concerned and happy in the presence of a happy or warm face (as opposed to a neutral or concerned face) and the presence of a happy or warm voice (as opposed to a neutral or concerned voice). Gong and Nass [29] tested participants’ responses to combinations of human versus humanoid (human-like but artificial) faces and voices using a talking-face agent. The pairing of a human face with a humanoid voice, or vice versa, led to less trust than the pairing of a face and a voice from either the human or the humanoid category.

2.3 Multisensory Interaction Research in HRI

With the exception of a handful of studies (e.g., [6,46,47]) that examine the perception of robotic facial expressions in the presence of incongruent contextual information (i.e., movie clips or pictures), the perception of multimodal emotional signals from robots has mainly focused on studies comparing responses to unimodal versus congruent bimodal emotional stimuli (e.g., [23–25]). These studies have examined the perception of emotional expressions involving faces and voices [23], faces and gestures [24], gestures and voices [25] and face-voice-gesture [23] combinations. The results follow a similar pattern to ECA studies. Recognition accuracy is higher, and attitudes towards robots are more favourable when participants observe congruent bi-modal expressions than unimodal expressions. For instance, Costa et al. [24] showed that congruent gestures are a valuable addition to the recognition of robot facial expressions. Another study [23] using speech, head-arm gestures, and facial expressions, showed that participants rated bimodal expressions consisting of head-arm gestures and speech as more clearly observable than the unimodal expressions consisting only of speech. Salem at al. [25] showed that a robot is evaluated more positively when hand and arm gestures are displayed alongside speech.

A small number of HRI studies (e.g., [6,46,47]) have examined the perception of robotic facial expressions in the presence of incongruent contextual information. The context was manipulated by having participants watch emotion-eliciting pictures or movie clips or listen to news clips with positive or negative emotional valence. Participants were then asked to rate the facial expressions of a robot (congruent vs. incongruent with the contextual valence). Overall, results showed that the recognition of robotic facial expressions is significantly better in the presence of congruent context, as opposed to no context or incongruent context. Providing incongruent context can even be worse than providing no context at all [47]. Furthermore, [46] showed that when the expressions of a robot are not appropriate given the context, subjects' judgments are more biased by the context than the expressions themselves. Finally, results also suggest that trust towards the robot is decreased when the robot's emotional response is incongruent with the affective state of the user [6]. The findings from the above-mentioned studies suggest that the recognition of robot emotional expressions and attitudes towards robots can be affected by a surrounding context, including the emotion-eliciting situation in which the expression occurs and the observer's emotional state [43].

Far less is known about the effects of mismatched or incongruous multimodal emotional information, especially when a robot conveys incongruous data from different channels. To the best of our knowledge, there are no HRI studies that investigate the experimental conflict situation where an observer receives incongruous information from a robot's body and voice, within the context of a social interaction scenario. Given that the voices and whole-body expressions of humanoid robots (such as NAO or Pepper) are increasingly used together to convey emotions, it is natural to question how incongruous emotional cues from these two modalities interact with each other, as well as with the contextual situation where the multimodal emotion occurs. Emotional expressions of humanoid robots are especially vulnerable to such conflicts, and artificial experimental conflicts produced in the laboratory can be seen as simulations of natural ones. Conflicts result from two main types of factors. One factor is related to the fact that synthetic modalities, such as the face and body of humanoids, typically contain only a few degrees of freedom, and synthetic speech is not yet ready to efficiently portray human-like emotions. Consequently, a conflict or mismatch may be created if, for example, a sad and empathic body expression is coupled with a monotone synthetic voice. Another source of conflict is noise, which usually affects one modality at a time (e.g., vision or audition). In these situations, the presented information may not adequately express an intended emotion, and it is the role of the observer to decide how to integrate incomplete or incongruous audio-visual information. As these examples highlight, conflict situations, where two sensory modalities receive incongruous information can easily occur in the context of social HRI, therefore, it is important to investigate human observers or robot interaction partners integrate different and incongruous emotional channels to arrive at emotional judgments about robots.

3 Materials and Methods

3.1 Experimental Design

We conducted a laboratory human-robot interaction experiment where participants were invited to watch movie clips, together with the humanoid robot Pepper. We manipulated the *socio-emotional context* of the interaction by asking participants to watch three emotion-eliciting (happiness, sadness, surprise) movie clips alongside the robot. Emotion elicitation using movie clips is a common experimental manipulation used in psychology studies of emotions [48], and has been successfully used in HRI studies to elicit emotional responses in healthy individuals in the laboratory (e.g. [47,49]). We also manipulated the *emotional congruence* of the multimodal reactions of the robot (consisting of vocal expressions and body postures) to each movie clip as follows (see Table 1):

– In the *congruent condition*, the emotional valence of the multimodal reaction of the robot was congruent with the valence of the socio-emotional context of the interaction (elicited by the movie clip). For example, the robot expresses a sadness in response to a sad movie clip.
– In the *contextually incongruous condition*, the emotional valence of the multimodal reaction of the robot was incongruous with the valence of the socio-emotional context of the interaction. For example, the robot expresses happiness in response to a sad movie clip).
– In the *cross-modally incongruous condition*, the multimodal reaction of the robot contains both congruent and incongruous cues with respect to the valence of the socio-emotional context of the interaction. For example, the robot expresses happy vocal expressions and sad body postures in response to a happy movie clip).

In the context of this study, (in)congruence is defined based on emotional valence (i.e., positivity/negativity of the emotion), according to the two dimensional categorical model of emotion proposed by Russel [50] (see Fig. 1). A number of previous studies have also suggested that the effects of congruence may vary by valence (e.g., [46,47,51]).

We chose to investigate the emotions of happiness, sadness, and surprise for a number of reasons. Firstly, happiness, sadness, and surprise are all "social emotions" [52], namely emotions that serve a social and interpersonal function in human interactions. This category of emotions is especially useful for social robots. Second, the expression of happiness, sadness, and surprise through body motion and vocal prosody has often been studied; thus by choosing these emotions, we were able to find reliable sources for the design of the audio-visual stimuli. Finally, we chose emotions which belong to different quadrants of the valence-arousal space

Table 1 The 3 × 3 experimental design with the two independent variables—*Socio-emotional context* of the interaction and *Emotional congruence* of the robot's reaction—resulting in 9 experimental conditions

Socio-emotional context	Congruent condition	Contextually incongruous condition	Cross-modally incongruous condition
Happy	Happy context/happy robot	Happy context/sad robot	Happy context/happy and sad robot
Sad	Sad context/sad robot	Sad context/happy robot	Sad context/happy and sad robot
Surprise	Surprise context/Surprised robot	Surprise context/sad robot	Surprise context/surprised and sad robot

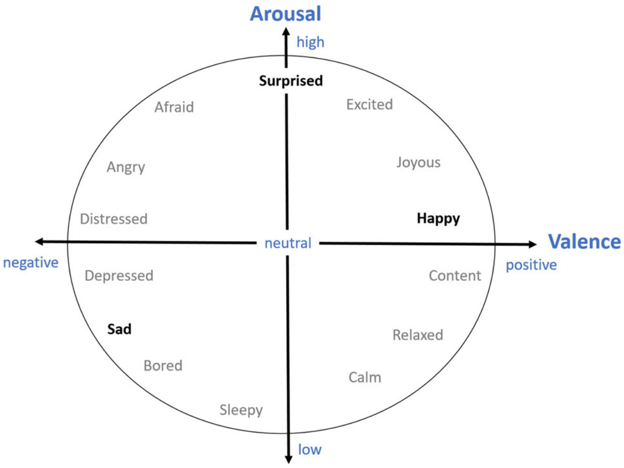

Fig. 1 The 2D valence-arousal model of emotion proposed by Russel [50]

[50]. As shown in see Fig. 1, happiness and surprise are both arousing emotions, which vary on only on the valence dimension. Happiness has positive valence while surprise can have any valence from positive to negative. Both these emotions contain clear action components in the body expression (in contrast to a sad body expression) [53]. In fact, body expressions of happiness and surprise share physical characteristics (i.e., large, fast movements and vertical extension of the arms above the shoulders) [53,54]. On the other hand, sadness and happiness differ in both valence and arousal; happiness has high arousal and positive valence, while sadness has low arousal and negative valence. Happiness and sadness share minimal body and vocal characteristics. To create prominent incongruous stimuli, we combined happiness with sadness and sadness with surprise.

We asked participants to label the emotional expressions of the robot and to rate the robot in terms of believability, perceived intelligence, and likability. In addition to these quantitative measures, we collected dispositional factors, namely the dispositional empathy of the participants. Empathy is defined as an affective response stemming from the understanding of another's emotional state or what the other person is feeling or would be expected to feel in a given situation [55]. It was included in the study since evidence suggests that individuals with a low level of dispositional empathy achieve lower accuracy in decoding facial expressions of humans [56] as well as emotional expressions of robots [12,57].

3.2 Participants

Participants were recruited through online and university advertisements. In total, 30 participants (mean = 29.5, SD = 4.82, 47% female, 53% male) who met the inclusion criteria (at least 18 years of age, basic English skills) were invited to the lab and completed the study. Participants gave informed consent and received monetary compensation for their participation (15 Euros).

3.3 Setting and Apparatus

The robot used in the study was Pepper by Softbank Robotics; a human-like robot with a full-motion body with 20 degrees of freedom. The experiment was carried out in a lab, furnished as a living-room environment, with a sofa, a small table with a laptop computer, and a large TV screen (see Fig. 2). The participants sat on the sofa, facing the TV screen, and the robot was placed between the participant and the TV, slightly to the right of the TV screen. Throughout the experimental session, the participant was observed via the built-in camera of the robot. A trained experimenter, in an adjacent room, utilized the video-feed to trigger the robot's emotional behaviour promptly.

3.4 Stimulus Material

3.4.1 Emotion Elicitation Movie Clips

Each participant watched three short emotion-eliciting movie clips, extracted from the following commercially available movies: An officer and a gentleman (happiness), The Champ (sadness), and Capricorn One (surprise). Target emotions and details about the movies are listed in Table 2. The procedure of validating the efficiency of these videos in eliciting the target emotions is discussed in Rottenberg et al. [57]. For a specific description of the scenes, see the Appendix of [57]. The order of presentation of the clips was randomized for each participant.

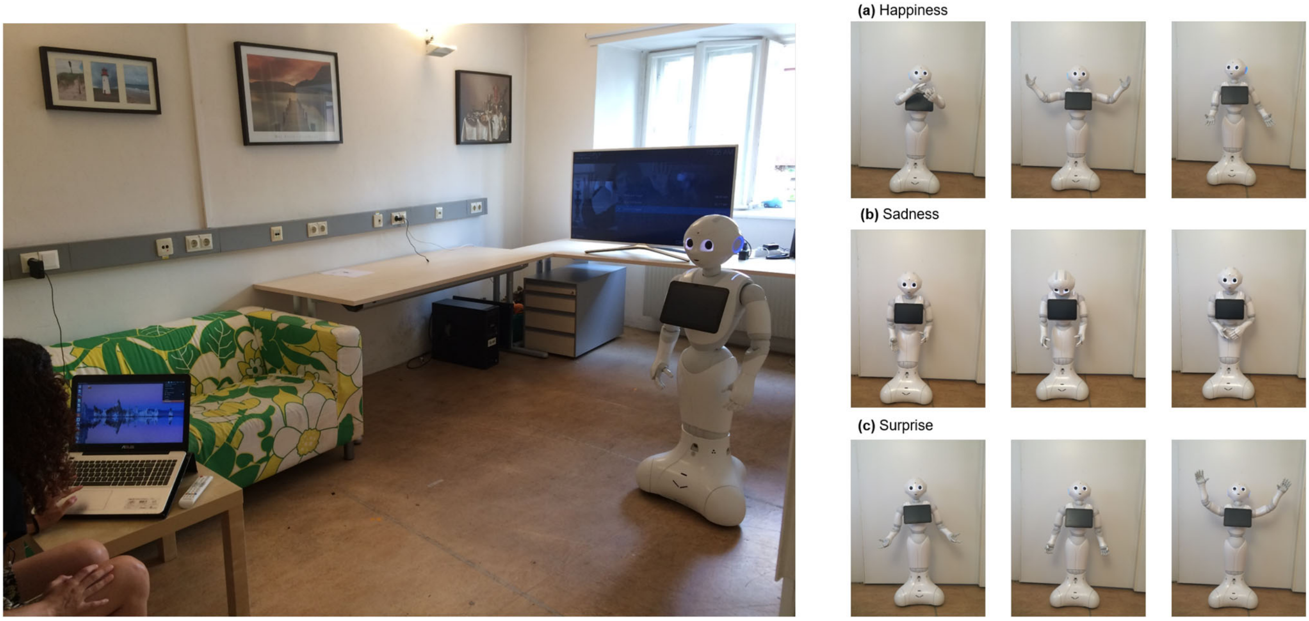

Fig. 2 (Left) Overview of the experimental setting. The robot is facing the participant to express an emotional reaction after a movie clip. (Right) Examples of body expressions for **a** happiness sequence (straight head, straight trunk, vertical and lateral extension of arms), **b** sadness sequence (forward head bent, forward chest bent, arms at side of trunk), and **c** surprise sequence (backward head bent, backward chest bent, vertical extension of arms)

Table 2 Target emotions and corresponding emotion-eliciting movies used in the experiment

Target emotion	Movie clip	Length (s)
Happiness	An officer and a gentleman	111
Sadness	The champ	171
Surprise	Capricorn one	49

3.4.2 Robot Emotional Expressions

In response to each movie clip, the robot expressed a multimodal emotional expression consisting of two modalities: auditory (vocal prosody), and visual (whole-body expression). The facial expression of the robot remained unchanged across all the stimuli (Pepper has a static face).

Pitch, timing, and loudness are the features of speech that are typically found to correlate with the expression of emotion through vocal prosody [17]. We manipulated these three features using the Acapela Text-to-Speech (TTS) engine (English language) to generate a set of vocal expressions. Our implementation was based on the phonetic descriptions of happiness, sadness, and surprise proposed by Crumpton et al. [17]. Given our interest in how vocal prosody (and not semantic information) influences emotion perception, the expressions were emotionally-inflected sentences with factual descriptions of the scenes shown in the movie clips, without any meaningful lexical-semantic cues suggesting the emotions of the robot (i.e., “A boxer is laying injured on the table and asks to see his son. A young boy approaches and starts talking to him”). No information regarding the age or gender of the robot could be derived from the speech. HRI research suggests that there is potential for non-linguistic utterances (NLUs) to be used in combination with language to mitigate any damage to the interaction should TTS generated language fail to perform at the desired level (e.g., [58]). In light of these findings, we decided to combine the sentences with a set of NLUs that emphasize the target emotion. NLUs were selected from an existing database of exemplars created in previous work [57] where evaluators rated each NLU using a forced-choice evaluation framework (Sadness, Happiness, Anger, Surprise, Neutral, I don’t Know and Other). All of the chosen NLUs were correctly recognized above chance level [57]. Table 3 summarizes the vocal prosody characteristics and NLUs we used for each target emotion. The resulting set was composed of three distinct sentence blocks (one for each video), each one recorded with different prosody features to portray two different emotions (congruent, incongruent with the situational valence). For example, for the “sadness” movie clip, the same sentence block was generated with two different prosody characteristics (pitch, timing, and loudness), and was combined with two different NLUs, for happiness and sadness respectively.

The vocal prosody stimuli were synchronized with (congruent and incongruent) body movements to create the audio-visual emotional expressions of the robot. The body

Table 3 Motion dynamics (velocity, amplitude), body animations (head, torso, arms) and vocal prosody characteristics (pitch, timing, loudness, non-linguistic utterances) used to generate audio-visual expressions for the target emotions

Target emotion	Motion dynamics	Head	Torso	Arms
Happiness	Large, fast movements	Straight head	Straight trunk	Vertical and lateral extension
Sadness	Small, slow movements	Forward head bent	Forward chest bent	Arms at side of the trunk
Surprise	Large, fast movements	Backward head bent	Backward chest bent	Vertical extension

Target emotion	Pitch	Timing	Loudness	NLU
Happiness	High	Moderate	High	Laughter/positive "Yay"
Sadness	Low	Small	Low	Cry/negative "oh"
Surprise	High	Large	Moderate	Gasp

expressions of the robot were modelled after the way humans move their head, torso, and arms to express emotions. Sources for our implementation were studies investigating the relevance of body posture and body movement features (i.e., velocity and amplitude) in conveying and discriminating between basic emotions in humans ([53,54,59,60]). De Silva and Bianchi-Berthouze [59] found that vertical features and features indicating the lateral opening of the body are informative for separating happiness from sadness. For instance, hands are raised and significantly more extended to indicate happiness and remain low along the body for sadness [59]. In a study by De Meijer [53] the trunk movement (ranging from stretching to bowing) was used to distinguish between positive and negative emotions. Based on these findings, in our design, happiness is characterized by a straight robot trunk, head bent back, a vertical and lateral extension of the arms and large, fast movements. Surprise is characterized by a straight trunk, backward stepping, and fast movements, whereas sadness, is characterized by a bowed trunk and head, downward and slow body movements. In a pre-evaluation online survey, we validated that people correctly perceived the emotion that each isolated body part is intended to convey [57]. We selected the animations that received the highest overall recognition score and used them to generate more complex animations for this study. Table 3 summarizes the whole-body expressions we used for each target emotion. Examples pf body expressions can be seen in Fig. 2.

3.5 Procedure

In order to avoid effects of expectation, participants were instructed that the experiment focuses on the ability of the robot to recognize emotions from audio-visual cues in the movie clips ("In this study, we test whether our robot can detect the emotional cues in the movie clips and can react accordingly"), instead of the actual aim. After reading a description of the experiment and signing a consent form, the participant was escorted to the lab where the experiment took place. Upon entering the room, the robot looked at the participant, waved and introduced itself ("Hello! I am Pepper. Welcome to the lab."). The experimenter then left the room, and the robot uttered, "We are going to watch some movies together! Start the first clip when you are ready", and turned towards the TV. While the participant watched a clip, the robot also looked at the TV. At the end of the clip, the robot turned towards the participant and expressed its emotional reaction to the clip (congruent, contextually incongruous or cross-modally incongruous). The duration of the robot's reactions varied between 20 and 50 s. During the rest of the time, the robot displayed idle movements (i.e., gaze/face tracking, breathing). We decided not to include any other type of verbal interaction between the robot and the participant, in order to minimize possible biasing effects on the participant's perception of the robot.

After the emotional reaction of the robot, an on-screen message on the laptop prompted the participant to answer an online questionnaire (built using the online tool Limesurvey) with questions about their experience of the movie clip and their perception of the robot's reaction (see Sect. 3.6). To limit carryover effects from one movie to the next, a 1-min rest period was enforced after completing the questionnaire. The participant was told to use this time to "clear your mind of all thoughts, feelings, and memories", before watching the next clip. This approach was originally used by Gross et al. [61] in their experiments on emotion elicitation using movie clips.

At the end of the third emotional expression of the robot, an on-screen message prompted the participant to answer a series of questions about demographics and personality traits (see Sect. 3.6). Afterwards, the robot thanked the participant and said goodbye ("Thank you for participating in this experiment. Goodbye!"). The experimenter then entered the room, answered any potential questions, debriefed the participant about the real purpose of the experiment and gave the monetary compensation. The experiment took about 60 min on average.

3.6 Measures

Manipulation Check—Experience of the Movie Clip To ascertain whether the desired emotion (happiness, sadness, surprise) had been properly elicited by the movie clip, we asked participants to report the most prominent emotion they experienced while watching the clip. Participants chose one option from a list of 11 emotions (amusement, anger, disgust, despair, embarrassment, fear, happiness/joy, neutral, sadness, shame, surprise) and the options neutral and other.

Emotion Recognition We asked participants to label the most prominent emotion expressed by the robot in response to each movie clip. Participants choose one option from a list of 11 emotions (amusement, anger, disgust, despair, embarrassment, fear, happiness/joy, neutral, sadness, shame, surprise) and the options neutral and other.

Attitudes Towards the Robot—Believability We asked participants to rate their perceptions about the believability of the robot. Participants rated seven conceptually distinct dimensions of believability (awareness, emotion understandability, behaviour understandability, personality, visual impact, predictability, behaviour appropriateness), as defined by Gomes et al. [62]. Table 4 contains the assertions used for each dimension. All items were rated on a 5-point Likert scale ranging from 1 "Strongly disagree" to 5 "Strongly agree" and "I don't know."

*Attitudes Towards the Robot—Perceived Intelligence and Likability*Participants rated the robot on the Perceived Intelligence and Likability dimensions of the Godspeed questionnaire [63]. All items were presented as a 5-point semantic differential scale.

Demographics and Personality Traits Participants reported basic socio-demographic information (age, gender, profession and previous experience with robots). Participants were also asked to fill in the Toronto Empathy Questionnaire [64], a 16-item self-assessment questionnaire assessing dispositional empathy.

Table 4 The seven items of the believability questionnaire, adopted by [62]

Believability item	Assertion
Awareness	The robot perceived the content of the movie clip correctly
Emotion understandability	It was easy to understand which emotion was expressed by the robot
Behaviour understandability	It was easy to understand what the robot was thinking about
Personality	The robot has a personality
Visual impact	The robots behaviour drew my attention
Predictability	The robots behaviour was predictable
Behaviour appropriateness	The behaviour expressed by the robot was appropriate for the content of the movie

3.7 Data Analysis

Manipulation Check—Experience of the Movie Clip Of the 90 movie clip ratings we obtained (30 participants × 3 clips per participant), 12 ratings were inconsistent with the intended situational valence manipulation (i.e., the movie clip failed to elicit the targeted emotion in the participant) and were thus excluded from further analyses. Consequently, the statistical analysis reported below was performed on the basis of a final sample of 78 ratings (Happy clip n = 25, Surprise clip n = 25, Sad clip n = 28).

Exploratory Regression Analysis As discussed in the previous sections, there are various factors that seem to influence the perception of robotic emotional expressions (i.e., the socio-emotional context of the interaction, incongruence between modalities, the rater's gender and dispositional empathy). Therefore, the first step of the data analysis was an exploratory regression analysis, performed to identify which factors would best account for whether or not participants correctly recognized the emotional expressions of the robot in this study. We coded the dependent variable (emotion recognition) as a binary value for whether or not the participant accurately recognized or not the expression of the robot and ran a logistic regression (a method typically used for such exploratory analyses [65]), to ascertain the effects of (in)congruence (congruent, contextually incongruous and cross-modally incongruous conditions), emotion being expressed (happiness, sadness and surprise), gender and dispositional empathy score on the likelihood that participants accurately recognize the emotional expression of the robot.

Emotion Recognition—Effects of Incongruence In the second step of the analysis, hit rate and unbiased hit rate [66] were analysed. These measures were chosen because we were interested in whether incongruence decreased target emotion detection rate. Hit rate (Hs) is the proportion of trials in which a particular emotion is shown that is correctly labelled. Although Hs is one of the most frequently used measures of accuracy, this metric does not take account false alarms (i.e., the number of times in which a particular emotion label is incorrectly used) or personal biases during the performance (i.e., the bias to say happy for all expressions). The unbiased hit rate (Hu), proposed by Wagner [66], takes this problem into account and results in calculations of accuracy rates that are more precise. The computation of Hu scores involves "the joint probability that a stimulus is correctly identified (given that it is presented) and that a response is correctly used (given that it is used)"(Wagner [66] p. 16). In other words, in order to measure recognition accuracy for a given emotion, the number of misses (e.g., the number of times in which a particular emotion was present and the participant responded it was absent) as well as the number of false alarms (e.g., the number of times in which the participant responded the target stimulus was present when in reality it wasn't) are taken into account. Hu scores were computed for each emotional expression of the robot as follows:

$$Hu = \frac{Ai}{Bi} \times \frac{Ai}{Ci}$$

where Ai = frequency of hits, Bi = number of trials where i is the target and Ci = frequency of i responses (hits and false alarms).
To investigate if people recognized the robot's emotional expressions correctly, we compared the emotion recognition ratings of the congruent and contextually incongruous conditions against the ideal distribution using a Chi-Square test. This statistical analysis approach was previously used in [11]. For instance, if ten people had to choose the right expression for the robot, out of a list of three different expressions (e.g., happy, sad, neutral), and the robot expressed "sadness", then the ideal distribution would be 0, 10, 0.

Next, to investigate the effects of *contextual incongruence* (i.e., the conflict situation where the robot's reaction is incongrous with the socio-emotional context of the interaction), we compared emotion recognition ratings of the congruent conditions (e.g., happy expression in response to a happy movie clip) against emotion recognition ratings of the contextually incongruent conditions (e.g., happy expression in response to a sad movie clip) by means of Chi-Square tests.

Thirdly, to investigate the effects of *cross-modal incongruence* (i.e., how incongruous auditory (vocal prosody), and visual (whole-body expression) cues are processed when presented simultaneously), we compared the emotion recognition ratings of the congruent condition (e.g., happy body and happy voice) against the emotion recognition ratings of the cross-modally incongruous condition (e.g., happy body and sad voice, sad body and surprised voice) by means of Chi-Square tests.

Attitudes Towards the Robot—Effects of Incongruence In the last part of the analysis, we investigated the effects of contextual incongruence and cross-modal incongruence on participants' attitudes towards the robot. The Believability, Perceived Intelligence, and Likability Questionnaires were calculated by summatively building up the scales. Cronbach's alpha was calculated to prove the internal reliability of the scales (all scales achieved a value higher than 0.7 and can thus be considered reliable). Since our data were not normally distributed, we used non-parametric tests, suitable for ordinal numerical data. Specifically, the Kruskal–Wallis H test and subsequent Mann–Whitney U tests were used to determine if there are statistically significant differences in the scores between the three experimental conditions (congruent, contextually incongruous, cross-modally incongruous).

Toronto Empathy Questionnaire The Toronto Empathy Questionnaire was calculated by reversing the inverted items and computing the summative score overall 16 items.

4 Results

4.1 Exploratory Regression Analysis

A logistic regression was performed to ascertain the effects of congruence (congruence, cross-modal incongruence, contextual incongruence), emotion being expressed (happiness, sadness, surprise), gender and dispositional empathy score on the likelihood that participants accurately recognize the emotional expression of the robot. The logistic regression model was statistically significant ($x^2(6)$ = 16.64, p = 0.01). The model explained 28.9% (Nagelkerke R^2) of the variance in emotion recognition accuracy and correctly classified 80.5% of cases. The results of the analysis are presented in Table 5 and discussed below.

Congruence There was a significant association (p = .01) between the congruent condition and the likelihood of correctly recognizing the emotional expression of the robot. Additionally, the cross-modally incongruous condition was significantly associated with a decreased likelihood of correctly recognizing the emotional expression of the robot (p = .01). There was no significant association between the contextually incongruous condition and the likelihood of correctly recognizing the emotional expression of the robot (p = .65). Nevertheless, the contextually incongruous condition was associated with a decreased likelihood of correctly recognizing the robot's emotional expression, compared to the congruent condition. An indication of the size of the effects of the contextually incongruous and the cross-modally incongruous conditions can be seen in the odds ratios reported in Table 5 (values discussed below as 1—odd ratio):

- The odds of people recognizing the emotional expression of the robot were 32% (or 1.47 times) lower in the contextually incongruous condition than in the congruent condition (baseline condition).
- The odds of people recognizing the emotional expression of the robot were 87% (or 7.69 times) lower in the cross-modally incongruous condition than in the congruent condition (baseline condition).

The effects of contextual congruence and cross-modal congruence on the likelihood that participants accurately recognized the emotional expression of the robot can also be seen

Table 5 Results of the regression analysis of congruence, emotion being expressed, gender and dispositional empathy score on the likelihood that participants accurately recognize the emotional expression of the robot

Variable	B	Odds ratio	Lower	95% CI for odds ratio Upper
Congruence: contextual incongruence[a]	– 0.37	0.68	0.13	3.63
Congruence: Cross-modal incongruence	– 1.97	0.13	0.03	0.62
Emotion: sadness[b]	– .86	0.42	0.93	1.91
Emotion: surprise	– .81	0.44	0.95	2.04
Gender: female[c]	– 1.23	0.29	0.08	1.03
Dispo. empathy Score	– 0.02	0.97	0.87	1.07
Constant	4.63	103.22		

R^2 = .54 (Hosmer and Lemeshow), 0.19 (Cox and Snell), 0.28 (Nagelkerke). Model $X^2(6) = 16.64$, $p = 0.01$.

[a] The coefficients for the contextually incongruous and cross-modally incongruous conditions are contrasts with the congruent condition (baseline).

[b] The coefficients for the emotional expressions of sadness and surprise are contrasts with the expression of happiness.

[c] The coefficients for gender are contrasts with male gender

in the hit rate (Hs) and unbiased hit rate (Hu) scores across the three conditions (see Table 7 and Sect. 4.2).

Emotion Being Communicated The results of the regression analysis showed no significant association between the type of emotion being communicated by the robot (i.e., happiness, sadness, surprise) and emotion recognition accuracy (p = .48). However, as indicated by the odds ratios (Table 5), the expressions of sadness were 57.8% (or 2.36 times) less likely to be recognized than expressions of happiness. Likewise, expressions of surprise were 55.9% (or 2.26 times) less likely to be recognized than expressions of happiness. As expected, due to its ambivalent nature, the emotion of surprise was the least well-recognized emotion in the congruent condition. However, the overall lowest recognition accuracy score was reported for the emotion of sadness, in the cross-modally incongruous condition (i.e., when the robot expressed sadness and happiness simultaneously).

Gender The results of the regression analysis showed a borderline level significance association between gender and emotion recognition accuracy (p =.05). Female participants were 3.43 (or 70.9%) times less likely to recognize the robot's emotion than males.

Dispositional Empathy The total score of the Toronto Empathy questionnaire can range from 0 (no empathy at all) to 64 (total empathy) [64]. Our analysis resulted in a mean value of 47.13 (SD = 6.41). There were no significant differences between female (mean = 49.42, SD = 5.44) and male (mean = 45.12, SD = 6.51) participants. Our results for female participants were slightly higher than the range given in the source of the questionnaire [64] (between 44 and 49 points). Likewise, our results for male participants were slightly higher than the source (between 43 and 45 points), indicating that, overall, our participants had a slightly higher dispositional empathy than average. The results of the regression analysis showed no significant association between empathy score and emotion recognition accuracy (p = .59). However, this analysis was performed after controlling for our manipulation check (i.e., participants who did not recognize the emotion elicited by the movie clip were excluded from this analysis). It is likely that those individuals in particular had a particularly low dispositional empathy score, and a regression analysis including their responses would reveal different results.

4.2 Emotion Recognition Accuracy: Effects of Incongruence

To provide all the relevant information about false alarms and potential biases in the emotion recognition ratings of participants, Table 6 provides a rapid overview of the detailed confusion matrices for the data. Table 7 summarizes the derived Hs and Hu scores across all experimental conditions and for the three target emotional expressions. Hu scores range between a minimum of zero to one, one indicating that all stimuli of an emotion have been correctly identified and the respective emotion has never been falsely chosen for a different emotion. We report the Hu scores, because of the popularity of this metric in the literature. However, we feel that this is not ideal for this study, because of the fact that some of our stimuli are not easily recognizable expression prototypes but rather complicated expressions, often based on the combination of mismatched emotional information across modalities. Therefore, in most cases, and especially in the cross-modally incongruous condition, the Hu severely reduces the hit rate for those responses that do not fit the target category, assuming that only one single answer can be correct"—namely, the one corresponding to the intended emotion category. For this reason, in the following subsections, we discuss the emotion recognition accuracy based on the Hs scores.

Table 6 Raw emotion recognition ratings, for each movie clip in the congruent (Cong.), cross-modally incongruous (Mod. Incong.) and contextually incongruous (Cont. Incong.) conditions

Emotion ratings	Happy clip (N = 25)			Sad movie clip(N = 28)			Surprise movie clip (N = 24)		
	Cong.	Mod. Incong.	Cont. Incong.	Cong.	Mod. Incong.	Cont. Incong.	Cong.	Mod. Incong.	Cont. Incong.
Amusement	0	1	0	0	1	3	1	0	0
Anger	0	0	0	0	0	0	0	0	0
Disgust	0	0	0	0	0	0	0	0	0
Despair	0	0	0	0	0	0	0	0	0
Embarrassment	0	0	0	0	0	0	0	0	0
Fear	0	0	0	0	0	0	0	0	0
Happiness/joy	**8**	**4**	1	0	**4**	**6**	0	0	0
Neutral	0	0	0	1	2	0	0	1	0
Sadness	0	3	**8**	**9**	2	0	0	0	**4**
Shame	0	0	0	0	0	0	1	0	0
Surprise	0	0	0	0	0	0	**6**	**7**	3
Other	0	0	0	0	0	0	0	1	1
N	8	8	9	10	9	9	8	9	7

Bold values indicate the emotion rated most frequently in every condition. Values with underline bold indicate the target emotion(s) expressed by the robot. N indicates the total number of ratings considered in the analysis

In the congruent condition, the mean Hs was 88% (SD= 0.13), with scores ranging from 75% to 100%, for the the three different emotions. In the contextually incongruous condition, the mean Hs was 61% (SD = 0.31), with scores ranging from as 29 to 89%. In the cross-modally incongruous condition, the mean Hs was 77% (SD = 0.26). In this particular condition, we consider that there is no correct target emotion since the robot simultaneously expresses two different emotions (i.e., happy body and sad voice). Therefore, separate Hs and Hs scores were calculated for each of the two target emotions expressed by the robot (see Table 7). For example, in the cross-modally incongruous condition where the robot simultaneously expressed happiness and sadness in response to a happy clip, 50% of the participants rated the expression as happiness, while 38% rated the expression as sadness (see Table 7). The remaining 12% corresponds to other emotion labels (see confusions in Table 6).

When participants watched a happy movie clip, and the robot expressed happiness (congruent condition), all the participants rated the expression of the robot as happy. However, when the robot expressed happiness in response to a sad clip (contextually incongruous condition), only 67% of the participants rated the expression of the robot as happy. The remaining 33% said that the robot was "amused" by the movie. When the robot expressed an audio-visual behaviour consisting both of happiness and sadness in response to a happy clip (cross-modally incongruous condition), 50% of the participants said that the robot was happy, while 37% found the robot to be sad.

When participants watched a sad movie clip, and the robot expressed sadness (congruent condition), 90% ra-ted the expression of the robot as sadness. When the robot expressed sadness in response to a happy clip (contextually incongruous condition), 89% rated the expression of the robot as sadness. In other words, sadness was well recognized both in a congruent context (i.e., context with similar emotional valence) and an incongruous context (i.e., context with opposing emotional valence).

To add more depth to our analysis, we also investigated what would happen if the robot expressed sadness in response to a clip eliciting the ambivalent emotion of surprise (contextually incongruous condition). In this case, 57% of the participants rated the robot's expression as sadness, while 43% rated the robot's expression as surprise. When the robot expressed an audio-visual behaviour consisting both of happiness and sadness in response to a sad clip (cross-modally incongruous condition), 22% of the participants said that the robot was sad, while 44% found the robot to be happy. Finally,when participants watched a movie clip electing the feeling of surprise, and the robot expressed a surprised reaction (congruent condition), 75% of the participants rated the robot's expression as surprise. Interestingly, when the robot expressed an audio-visual behaviour consisting both of sur-

Table 7 Measures of accuracy across the three experimental conditions. Stimulus hit rate (Hs) and unbiased hit rate (Hu)

Experimental condition	Movie clip	Target emotion	Hs (raw)	Hs (%)	Hu
Congruent	Happy	Happiness	1.00	100	1.00
	Sad	Sadness	0.90	90	0.90
	Surprise	Surprise	0.75	75	0.75
Contextually incongruous	Happy	Sadness	0.89	89	0.59
	Sad	Happiness	0.67	67	0.57
	Surprise	Sadness	0.29	29	0.29
Cross-modally incongruous	Happy	Happiness (+ sadness)	0.50	50	0.25
	Happy	Sadness (+ happiness)	0.38	38	0.23
	Sad	Sadness (+ happiness)	0.22	22	0.09
	Sad	Happiness (+ sadness)	0.44	44	0.22
	Surprise	Surprise (+ sadness)	0.78	78	0.78
	Surprise	Sadness (+ surprise)	0.00	0	0.00

Table 8 Chi-square tests on emotion recognition ratings for the emotional expressions (EE) of happiness, sadness and surprise. Figures in bold show non-significant results ($p > .01$), meaning that the ratings did not differ significantly between the two conditions being compared

Comparison	Happiness EE	Sadness EE	Surprise EE
Congruent versus ideal distribution	Context happy X^2 **(11, N = 8) = .00, *p*** = 1.00	Context Sad X^2 **(11, N = 10) = .10 , *p*** = 1.00	Context surprise X^2 **(11, N = 8) = 8.00, *p*** = .71
Contextually incongruous versus ideal distribution	Context sad X^2 **(11, N = 9) = 1.00, *p*** = 1.00	Context happy X^2 **(11, N = 9) = .11 , *p*** = 1.00 Context surprise X^2 **(11, N = 7) =1.29 , *p*** = 1.00	N/A
Congruent versus contextually incongruous	Context happy versus context sad X^2 **(11, N = 8/N = 9) = 3.66, *p*** = .98	Context sad versus context happy $X^2 (11, N = 10/N = 9) = NA, p = .00$ Context sad versus context surprise $X^2 (11, N = 10/N = 7) = NA, p = .00$	N/A
Congruent versus cross-modally incongruous	Context happy X^2**(11, N=8) = 8.00, *p*** = .71	Context sad X^2 (11, N=10/N=9) = 30.00, p = .00	Context surprise $X^2 (11, N = 9) = NA, p = .00$

prise and sadness (cross-modally incongruous condition), 78% of the participants said that the robot was surprised, while no one rated the expression of the robot as sad.

Table 8 summarizes the results of the Chi-Square tests. The tests comparing the emotion recognition ratings of the congruent condition against the ideal distribution were not significant for all the emotions. In other words, there was no significant difference between the emotion recognition ratings of the participants and the ideal distribution (which does not mean that the result equals the ideal distribution, but only that it did not significantly differ from it). Likewise, the Chi-Square tests comparing the emotion recognition ratings for the contextually incongruous condition against the ideal distribution were not significant for all the emotions. These results indicate that the participants were able to recognize the emotional expressions of the robot, both in the baseline congruent condition, but also when the expressions were incongruous with the emotional valence of the interaction context. These findings are in line with the results of the regression analysis (i.e., there was no significant association between the contextually incongruous condition and the likelihood of correctly recognizing the emotional expression of the robot). To investigate further the effects of contextual incongruence on the likelihood of correctly recognizing the emotional expression of the robot, we compared the emotion recognition ratings of the participants in the congruent condition (i.e., happy emotional expression in response to a happy movie clip) and contextually incongruous condition (i.e., happy emotional expression in response to a sad movie clip) using Chi-Square tests. The result was not significant (X^2 (11, N = 8/N = 9) = 3.66, $p = .98$). In other words, there was no significant difference between the emotion recognition ratings of the participants in these two conditions. However, there was a significant difference in the ratings for the emotional expression of sadness in the congruent condition, compared to the contextually incongruous condition where the robot expressed sadness in response to a movie clip eliciting happiness (X^2 (11, N = 10/N = 9) = NA, $p = .00$); as well as the contextually incongruous condition where the robot expressed sadness in response to a movie clip eliciting surprise (X^2 (11, N = 10/N = 7) = NA, $p = .00$). The regression analysis revealed a negative association ($p = .01$) between the cross-modally incongruous condition and the likelihood of correctly recognizing the emotional expression of the robot. To investigate this finding further, we compared the emotion recognition ratings of the congruent condition and cross-modally incongruous condition using Chi-Square tests. There was no significant effect of cross-modal congruence on the ratings of happiness (X^2 (11, N = 8) = 8.00, $p = .71$). In other words, the ratings of the participants for the expression of happiness did not differ significantly between the congruent and cross-modally incongruous conditions. However, there was a significant effect on the ratings of sadness (X^2 (11, N = 10/N = 9) = 30.00, $p = .0016$) and surprise (X^2 (11, N = 9) = NA, $p = .00$). In other words, the emotion ratings of the participants differed significantly between the congruent and cross-modally incongruous conditions.

4.3 Attitudes Towards the Robot: Effects of Incongruence

Believability A Kruskal–Wallis H test showed a statistically significant difference in the believability scores (mean of the seven distinct dimensions) between the three different experimental conditions (congruent, contextually incongruous, cross-modally incongruous), H (2) = 21.16, $p < 0.001$, with a mean rank recognition score of 54.33 for the congruent condition, 33.04 for the cross-modally incongruous condition and 27.50 for the contextually incongruous condition (Fig. 3). Subsequent MannWhitney tests, were used to make post-hoc comparisons between conditions. The believability scores for the congruent condition were significantly different from the contextually incongruous condition (U = 91.00, $p < .001$). Also, the believability scores were significantly higher for the congruent condition (Mdn = 35.00) than the contextually incongruous condition (Mdn = 16.64). The believability scores for the congruent condition were significantly different from the cross-modally incongruous condition (U = 145.50, $p < .001$). More specifically, the scores were significantly higher for the congruent condition (Mdn = 32.83) than for the cross-modally incongruous condition (Mdn = 18.90). There was no significant difference between the contextually incongruous and the cross-modally incongruous conditions (U = 271.50, $p = .42$).

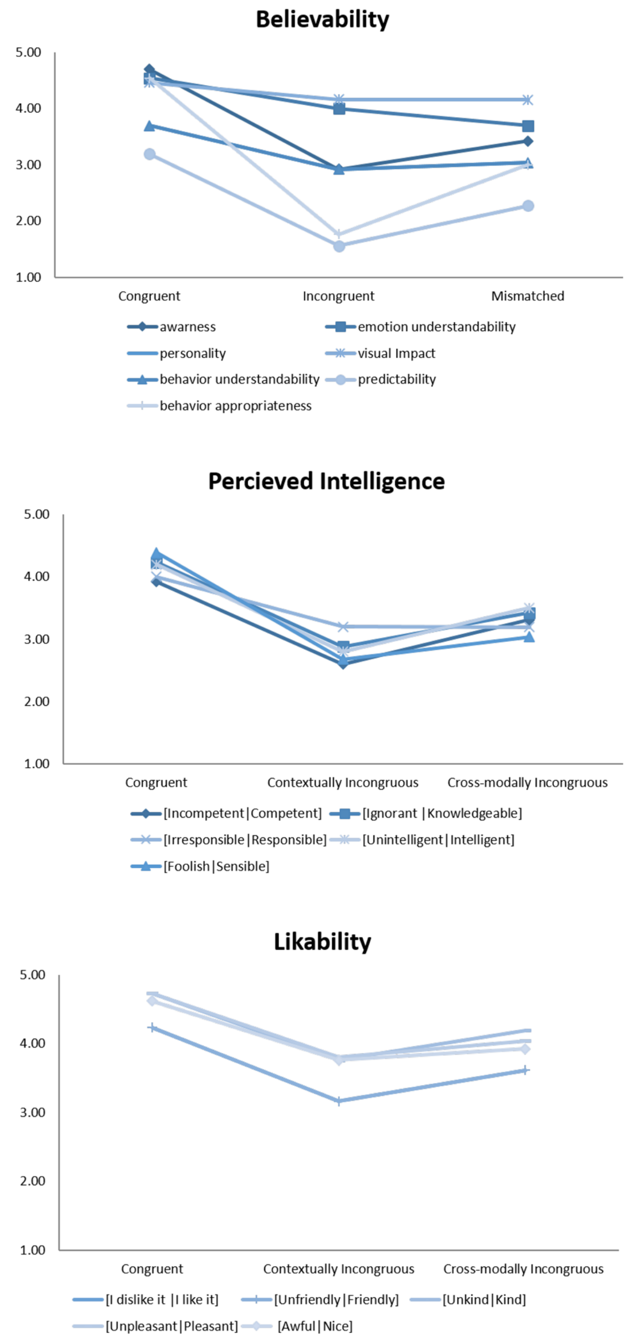

Fig. 3 Mean ratings for Believability (7 items) and Godspeed Questionnaire items: Perceived Intelligence (5 items) and Likability (5 items)

Perceived Intelligence A Kruskal–Wallis H test showed a statistically significant difference in the perceived intelligence score between the three different conditions, H (2) = 20.46, $p < 0.001$, with a mean rank recognition score of 54.23 for the congruent condition, 35.75 for the cross-modally incongruous condition and 26.54 for the contextually incongruous condition (Fig. 3). Subsequent MannWhitney tests, showed that the perceived intelligence scores for the congruent condition were significantly different from the contextually incongruous condition (U = 103.50, p <.001) and the cross-modally incongruous condition (U = 163.500, p <.005). There was no significant difference between the scores for the contextually incongruous and cross-modally incongruous conditions (U = 235.00, p =.08). The perceived intelligence score was significantly higher for the congruent condition (Mdn = 34.52) than for the contextually incongruous condition, (Mdn = 17.14). Additionally, the perceived intelligence score was significantly higher for the congruent condition (Mdn = 33.21) than for the cross-modally incongruous condition (Mdn = 19.79).

Likability A Kruskal–Wallis H test showed that there was a statistically significant difference in the likability score between the three different context conditions, H (2) = 8.25, p < 0.05, with a mean rank recognition score of 48.69 for the congruent condition, 31.30 for the cross-modally incongruous condition and 36.71 for the contextually incongruous condition (Fig. 3). Subsequent MannWhitney tests, showed that the congruent condition was significantly different from the contextually incongruous condition (U = 180.50, p <.01) and the cross-modally incongruous condition (U = 230.500, p < .005). There was no difference between the scores for the contextually incongruous and the cross-modally incongruous conditions (U = 277.00, p = .36). The likability score was significantly higher for the congruent condition (Mdn = 31.56) than for the contextually incongruous condition, (Mdn = 20.22). Additionally, the likability score was significantly higher for the congruent condition (Mdn = 30.63) than for the cross-modally incongruous condition (Mdn = 22.37).

5 Discussion

5.1 Overview and Significance of Results

An exploratory regression analysis showed that cross-modal incongruence (i.e., the conflict situation where an observer receives incongruous emotional information across the auditory (vocal prosody) and visual (whole-body expressions) modalities) is associated with a decreased likelihood of correctly recognizing the emotional expression of the robot. In addition, both cross-modal incongruence and contextual incongruence (i.e., the conflict situation where the robot's reaction is incongrous with the socio-emotional context of the interaction) negatively influence attitudes towards the robot in terms of believability, perceived intelligence and likability. The significance of these findings is discussed in more detail below, and in Sect. 6 we provide a number of recommendations regarding design choices for humanoid robots that use several channels to communicate their emotional states in a clear and effective way.

Effects of Cross-modal Incongruence Interesting findings were obtained regarding the effects of cross-modal incongruence. Statistically significant results indicate that when emotional information is incongruous across the auditory and visual modalities, the likelihood that people accurately recognize the emotional expression of the robot is significantly decreased, compared to the situation where congruent information is presented across the two channels. Our findings are in line with previous work in psychology, neuroscience and HCI literature and suggest that theories about MI in Human–Human interactions (e.g., [34–45]) and Human–Agent interactions (e.g., [26–30]), also extend to Human–Robot interactions. The descriptive analysis of the emotion recognition scores revealed that emotional expressions that contained a happy body and a sad voice (or vice versa) resulted in a confused perception, where the emotional expression of the robot was perceived by some people as happiness and by others as sadness. A few people also labeled one of the incongruous expressions as neutral (see confusions in Table 6). This suggests that neither the visual nor the auditory channel dominated the participants' perception, and therefore, the responses of the participants were split between the two emotions expressed by the robot. Regarding those who rated the expression neither as happiness nor as sadness, it is possible that they tried to make the robot's emotional expression consistent on both channels, by rating it as something in between happy and sad (i.e., neutral or amused). On the other hand, when participants watched the robot expressing an incongruous multimodal combination of surprise and sadness, the highly expressive body and voice cues of surprise (i.e., large, fast movements, extension of the arms, gasping utterance) dominated over the more subtle cues of sadness (i.e., small, slow movements, arms at side of trunk), and as a result no one rated the expression of the robot as sadness (see Table 7).

Effects of Contextual Incongruence Previous work indicates the contextual effect on the recognition of robot [46,47] and human [45] emotional expressions. For instance, Niedenthal et al. [43] found that observers may be influenced by their own emotional state when they were asked to attribute emotional states to human facial expressions. In the first phase of our analysis, we found no significant association between incongruous context and the likelihood of correctly recognizing the emotional expression of the robot. Nevertheless, a more in-depth analysis of the emotion recognition ratings of the participants, showed that the recognition accuracy scores for all emotions were lower in the presence of incongruous socio-emotional context, compared to congruent context. The effects of contextual incongruence on the likelihood that participants accurately recognize the emotional expression of the robot were most prominent in the case of surprise, an ambivalent emotion which can potentially result in ambiguous emotion recognition ratings. Specifically, when the robot expressed sadness in response to a movie clip eliciting the emotion of surprise, we observed confused assessments about the robot's emotional expression and a significant drop in the recognition accuracy scores. When we took our analysis a step further and compared the participant's ratings for each emotion in the congruent and contextually incongruous conditions, statistically significant results showed that the recognition accuracy scores for the sadness and surprise expressions differed between these two conditions. For the case of happiness, we did not find a significant difference between the ratings of the participants in the two conditions. This finding does not undermine the importance of the context but instead suggests that, in our study, expressions of happiness had a more dominant effect than the socio-emotional context (elicited by the movie clip) on the emotion recognition ratings of the participants. This can be explained by the fact that happiness is a basic emotion which is, in general, easy to recognize.

Impact on Perceived Believability, Intelligence, and Likability With regard to the effects of incongruence on people's attitudes towards the robot, we found that both contextual incongruence, as well as cross-modal incongruence, significantly reduced participants' ratings of believability, likability and perceived intelligence of the robot. For instance, in these two conditions, the robot was rated as less intelligent, less responsible and less kind than the congruent condition. Furthermore, since there was no statistically significant difference between the believability and intelligence ratings in the contextually incongruous and the cross-modally incongruous conditions, we can conclude that the conflict situation where information conveyed in one of the communication modalities (auditory or visual) is incongruous with the socio-emotional context of the interaction is almost as harmful as the conflict situation where the robot's overall mutlimodal reaction is incongrous with the socio-emotional context of the interaction.

5.2 Limitations

The findings reported above should be considered in light of possible limitations and constraints of our laboratory experiment. There are obviously numerous robot-related and person-related factors that could influence the perception of the robot and its emotional expressions. However, it is practically and methodologically impossible to investigate and control them all within a single study.

In terms of robot-related factors, we limited our research to audio-visual expressions based on the body and voice of the robot. The results partially indicate that the choice of a robot without facial expressions, and the selected emotions made it hard for participants to distinguish between certain emotions (e.g., the ambivalent emotion of surprise). Moreover, another limitation is the fact that our robot had a static face, which to some extent reflected "happiness", due to the arrangement of the eyes and mouth. It is undeniable that the face of the robot was yet another affective signal, which participants integrated with the body and voice signals, to arrive at emotional judgments. Although we tried to avoid the communication of emotions through the linguistic content of the robot's speech, it is likely that this was another factor that influenced participant's judgments, especially since the robot was communicating via natural language [17]. In terms of person-related factors that influence the perception of a robot and its emotional expressions, the design of this study did not consider cultural specificity. There are numerous theories about the role of culture in human–human [67,68] and human–robot [69] affective interaction. However, most of the participants in our study were Austrian, meaning that the results may not be directly transferable to participants with a different cultural background. Future studies should consider a larger sample size and participants with more diverse characteristics. Moreover, this study focused only on a small set of three emotions (happiness, sadness, and surprise). Future studies should include more emotions, such as Ekman's basic emotions [70], or more complex emotions (e.g., embarrassment) which are more difficult to simulate in a robot. Lastly, although we have used previously validated movies (see [61] for details) for our emotion elicitation manipulation, it remains unclear whether the different length of the chosen movie-clips (see Table 2) had an impact on the elicitation of the target socio-emotional context of the human–robot interaction.

These limitations do not necessarily mean that the results of this work cannot be generalized. The integration of incongruent voice and body signals has not been empirically investigated with humanoid robots prior to this research. Thus, our work provides important insight for researchers and designers of humanoid social robots. The primary findings of our experiment suggest that incongruent body and voice emotional cues result in confused perceptions of emotion and influence attitudes towards a robot in a negative way. It is likely that investigating different incongruent visual and auditory emotional cues, with other robots, in different interaction contexts, will also reveal interesting information towards understanding the perception of multimodal emotional expressions of social robots.

6 Conclusion

This article discussed how incongruous emotional signals from the body and voice of humanoid robot influence people's ability to identify and interpret a robot's emotional state, and how incongruence impacts attitudes towards a robot during socio-emotional interactions. Our laboratory HRI study showed that incongruous audio-visual expressions do not only impair emotion recognition accuracy, but also decrease the perceived intelligence of a robot, making its behaviour seem less believable and less likable. These findings highlight the importance of properly designing and evaluating multimodal emotional expressions for robots intended to interact with people in the context of real socio-emotional HRI environments.

Based on our findings, we provide some recommendations regarding design choices for robots that use several channels to communicate their emotional states in a clear and effective way. When emotional information is incongruous across the auditory and visual channels, the likelihood that people accurately recognize the emotional expression of the robot significantly decreased. Therefore, great attention to detail is required when attempting to simulate multimodal emotional expressions. Designers must ensure that the different channels used to express emotions are appropriate and congruent with each other. Conflict situations, where two sensory modalities receive incongruous information can easily occur in the context of social HRI in real-world environments. For example, humanlike robots use highly expressive body postures or facial expressions to convey emotion, but often combine those with synthetic voices which lack the naturalness of human voice. If designers are able to anticipate the channel upon which a user will rely on when making emotional assessments about a robot, then they can tailor the information presented in that channel, to maximize the recognizability of the robot's expression. Conflict situations can also occur due to noise, which usually affects only one modality (i.e., sight or hearing). For instance, if the environment is very loud (e.g., a school or hospital), the voice of the robot may be masked, and the resulting audio-visual expression may not adequately convey the intended emotion. Hence, when choosing communication modalities for a robot, a designer should consider information about the environment.

Given the fact that contextual incongruence (i.e., a robot expresses happiness in response to a sad situation) can have a detrimental effect on the believability, likability and perceived intelligence of the robot, as a general guideline, designers should not only assess whether the multimodal emotional expressions of a robot are accurately recognized, but also which effects they have on the interaction partners' attitudes towards the robot. Furthermore, in certain social contexts, if a robot's perception capabilities are not precise enough, designers may need to reconsider the use of emotional expressions. In other words, if a robot is likely to make a mistake in assessing the affective state of the user or the socio-emotional context of the interaction, then it might be better to opt for neutrality instead of showing an emotional reaction that is inappropriate. This can result in the robot appearing unintelligent, irresponsible or even unkind towards its human interaction partner.

Acknowledgements Open access funding provided by TU Wien (TUW). The authors wish to thank the study participants for their collaboration, and the members of the Vision4Robotics group at TUW for their assistance in setting up the study.

Funding This study was funded by the Swiss National Science Foundation (Grant No. P1GEP2-168609) and the European research projects H2020 GrowMeUp (Grant No. 643647), H2020 WellCo (Grant No. 769765), AAL CoME (Grant No. AAL-2014-7-127), COST AAPELE (Grant No. IC1303).

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. Fong T, Nourbakhsh I, Dautenhahn K (2003) A survey of socially interactive robots. Robot Auton Syst 42(3–4):143–166
2. Hall J, Tritton T, Rowe A, Pipe A, Melhuish C, Leonards U (2014) Perception of own and robot engagement in humanrobot interactions and their dependence on robotics knowledge. Robot Auton Syst 62(3):392–399
3. Eyssel F, Hegel F, Horstmann G, Wagner C (2010) Anthropomorphic inferences from emotional nonverbal cues: a case study. In: 19th international symposium in robot and human interactive communication, pp 646–651
4. Sidner CL, Lee C, Kidd CD, Lesh N, Rich C (2005) Explorations in engagement for humans and robots. Artif Intell 166(12):140–164
5. Breazeal C (2003) Emotion and sociable humanoid ro- bots. Int J Hum-Comput Stud 59(12):119–155
6. Cramer H, Goddijn J, Wielinga B, Evers V (2010) Effects of (in)accurate empathy and situational valence on attitudes towards robots. In: 2010 5th ACM/IEEE international conference on human–robot interaction (HRI), pp 141–142
7. Goetz J, Kiesler S, Powers A (2003) Matching robot appearance and behavior to tasks to improve human–robot cooperation. In: The 12th IEEE international workshop on robot and human interactive communication, proceedings. ROMAN 2003, pp 55–60
8. Partan S, Marler P (1999) Communication goes multimodal. Science 283(5406):1272–3
9. Pantic M, Rothkrantz LJM (2000) Automatic analysis of facial expressions: the state of the art. IEEE Trans. Pattern Anal. Mach. Intell. 22(12):1424–1445
10. Chandrasekaran B, Conrad JM (2015) Human–robot collaboration: a survey. SoutheastCon 2015:1–8
11. Mavridis N (2015) A review of verbal and non-verbal humanrobot interactive communication. Robot Auton Syst 63:22–35
12. Mirnig N, Strasser E, Weiss A, Khnlenz B, Wollherr D, Tscheligi M (2014) Can you read my face? Int J Soc Robot 7(1):63–76

13. Canamero L, Fredslund J (2001) I show you how i like you—can you read it in my face? Robotics. IEEE Trans Syst Man Cybern Part A Syst Hum 31(5):454–459
14. Lazzeri N, Mazzei D, Greco A, Rotesi A, Lanat A, De Rossi DE (2015) Can a humanoid face be expressive? A psychophysiological investigation. Front Bioeng Biotechnol 3:64
15. Bennett CC, Sabanovic S (2014) Deriving minimal features for human-like facial expressions in robotic faces. Int J Soc Robot 6(3):367–381
16. Bennett C, Sabanovic S (2013) Perceptions of Affective Expression in a minimalist robotic face. In: 2013 8th ACM/IEEE international conference on human–robot interaction (HRI), pp 81–82
17. Crumpton J, Bethel CL (2016) A survey of using vocal prosody to convey emotion in robot speech. Int J Soc Robot 8(2):271–285
18. Zecca M et al (2009) Whole body emotion expressions for KOBIAN humanoid robot preliminary experiments with different emotional patterns. In: RO-MAN 2009—the 18th IEEE International symposium on robot and human interactive communication, 2009, pp 381–386
19. Beck A, Canamero L, Bard KA (Sep. 2010) Towards an affect space for robots to display emotional body language. In: 19th IEEE international symposium on robot and human interactive communication principe, pp 464–469
20. McColl D, Nejat G (2014) Recognizing emotional body language displayed by a human-like social robot. Int J Soc Robot 6(2):261–280
21. Knight H, Simmons R (2016) Laban head-motions convey robot state: a call for robot body language. In: 2016 IEEE international conference on robotics and automation (ICRA), 2016, pp 2881–2888
22. Hortensius R, Hekele F, Cross ES (2018) The perception of emotion in artificial agents. IEEE Trans Cognit Dev Syst 10(4):70
23. Aly A, Tapus A (2015) Multimodal adapted robot behavior synthesis within a narrative human–robot interaction. In: 2015 IEEE/RSJ International Conference on Intelligent robots and systems (IROS), 2015 pp 2986–2993
24. Costa S, Soares F, Santos C (2013) Facial expressions and gestures to convey emotions with a humanoid robot. In: Social robotics, vol 8239. Springer, pp 542–551
25. Salem M, Rohlfing K, Kopp S, Joublin F (2011) A friendly gesture: investigating the effect of multimodal robot behavior in human–robot interaction. In: 2011 RO-MAN 2011, pp 247–252
26. Clavel C, Plessier J, Martin J-C, Ach L, Morel B (2009) Combining facial and postural expressions of emotions in a virtual character. In: Proceedings of the 9th international conference on intelligent virtual agents. Springer, pp 287–300
27. Creed C, Beale R (2008) Psychological responses to simulated displays of mismatched emotional expressions. Interact Comput 20(2):225–239
28. Mower E, Mataric MJ, Narayanan S (2009) Human perception of audio-visual synthetic character emotion expression in the presence of ambiguous and conflicting information. IEEE Trans Multimed 11(5):843–855
29. Gong L, Nass C (2007) When a talking-face computer agent is half-human and half-humanoid: human identity and consistency preference. Hum Commun Res 33:163–193
30. Becker C, Prendinger H, Ishizuka M, Wa-chsmuth I (2005) Evaluating affective feedback of the 3D agent max in a competitive cards game. Springer, Heidelberg, pp 466–473
31. Godfroy-Cooper M, Sandor PMB, Miller JD, Welch RB (2015) The interaction of vision and audition in two-dimensional space. Front Neurosci 9:311
32. de Gelder B, Vroomen J, Pourtois G (2004) Multisensory perception of emotion, its time course and its neural basis. In: The handbook of multisensory processes
33. De Gelder B, Bertelson P (2003) Multisensory integration, perception and ecological validity. Trends Cognit Sci 7(10):460–467
34. de Gelder B, Vroomen J (2000) The perception of emotions by ear and by eye. Cognit Emot 14(3):289–311
35. Collignon O et al (2008) Audio-visual integration of emotion expression. Brain Res 1242:126–135
36. Kreifelts B, Ethofer T, Grodd W, Erb M, Wildgruber D (2007) Audiovisual integration of emotional signals in voice and face: an event-related fMRI study. In: NeuroImage
37. Meeren HKM, van Heijnsbergen CCRJ, de Gelder B (2005) Rapid perceptual integration of facial expression and emotional body language. Proc Nal Acad Sci USA 102(45):16518–23
38. Van den Stock J, Righart R, de Gelder B (2007) Body expressions influence recognition of emotions in the face and voice. Emotion 7(3):487494
39. Stienen BMC, Tanaka A, de Gelder B (2011) Emotional voice and emotional body postures influence each other independently of visual awareness. PLoS ONE 6(10):e25517
40. Vines BW, Krumhansl CL, Wanderley MM, Levitin DJ (2006) Cross-modal interactions in the perception of musical performance. Cognition 101:80–113
41. Mahani M-AN, Sheybani S, Bausenhart KM, Ulrich R, Ahmadabadi MN (2017) Multisensory perception of contradictory information in an environment of varying reliability: evidence for conscious perception and optimal causal inference. Sci Rep 7(1):3167
42. Mobbs D, Weiskopf N, Lau HC, Featherstone E, Dolan RJ, Frith CD (2006) The Kuleshov effect: the influence of contextual framing on emotional attributions. Soc Cognit Affect Neurosci 1:95–106
43. Niedenthal PM, Kruth-Gruber S, Ric F (2006) What information determines the recognition of emotion? Principles of social psychology, pp 136–144
44. Carroll JM, Russell JA (1996) Do facial expressions signal specific emotions? Judging emotion from the face in context. J Personal Soc Ppsychol 70(2):205–18
45. Feldman Barrett L, Mesquita B, Gendron M (2012) Context in emotion perception. Curr Dir Psychol Sci 20(5):286–290
46. Zhang J, Sharkey AJC (2012) Its not all written on the robots face. Robot Auton Syst 60(11):1449–1456
47. Bennett CC, Sabanovic S, Fraune MR, Shaw K (2014) Context congruency and robotic facial expressions: do effects on human perceptions vary across culture? In: The 23rd IEEE international symposium on robot and human interactive communication, pp 465–470
48. Kreibig SD (2010) Autonomic nervous system activity in emotion: a review. Biol Psychol 84(3):394–421
49. Aly A (2015) Towards an interactive human–robot relationship: developing a customized robot behavior to human profile. Doctoral dissertation, ENSTA ParisTech
50. Russell JA (1980) A circumplex model of affect. J Personal Soc Psychol 39(6):1161–1178
51. Aviezer H et al (2008) Angry, disgusted, or afraid? Psychol Sci 19(7):724732
52. Hareli S, Parkinson B (2008) Whats social about social emotions? J Theory Soc Behav 38(2):131–156
53. Kleinsmith A, Bianchi-Berthouze N (2013) Affective body expression perception and recognition: a survey. IEEE Trans Affect Comput 4(1):15–33
54. Coulson M (2004) Attributing emotion to static body postures: recognition accuracy, confusions, and viewpoint dependence. J Nonverbal Behav 28(2):117–139
55. Eisenberg N, Shea C, Carlo G (2014) “Empathy-related responding and cognition:a chicken and the egg dilemma. Handb Moral Behav Dev 2:63–68
56. Dapretto M et al (2006) Understanding emotions in others: mirror neuron dysfunction in children with au- tism spectrum disorders. Nat Neurosci 9(1):28–30

57. Tsiourti C, Weiss A, Wac K, Vincze M (2017) Designing emotionally expressive robots: a comparative study on the perception of communication modalities. In: Proceedings of the 5th international conference on human agent interaction (HAI 2017)
58. Read R, Belpaeme T (2014) Non-linguistic utterances should be used alongside language, rather than on their own or as a replacement. In: Proceedings of the 2014 ACM/IEEE international conference on Human–robot interaction—HRI 14, pp 276–277
59. De Silva PR, Bianchi-Berthouze N (2004) Modeling human affective postures: an information theoretic characterization of posture features. Comput Anim Virtual Worlds 15(34):269–276
60. de Meijer M (1989) The contribution of general features of body movement to the attribution of emotions. J Nonverbal Behav 13(4):247–268
61. Gross JJ, Levenson RW (1995) Emotion elicitation using films. Cognit Emot 9(1):87–108
62. Gomes P, Paiva A, Martinho C, Jhala A (2013) Metrics for character believability in interactive narrative. In: Koenitz H, Sezen TI, Ferri G, Haahr M, Sezen D, Catak G (eds) Interactive storytelling, vol 8230. Springer, Cham, pp 223–228
63. Bartneck C, Kulic D, Croft E, Zoghbi S (2009) Measurement Instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. Int J Soc Robot 1(1):71–81
64. Spreng RN, McKinnon MC, Mar RA, Levine B (2009) The Toronto empathy questionnaire: scale development and initial validation of a factor-analytic solution to multiple empathy measures. J Personal Assess 91(1):62–71
65. Howell (2009) Statistical methods for psychology. Cengage Learning
66. Wagner L (1993) On measuring performance in category judgment studies of nonverbal behavior. J Nonverbal Behav 17(1):3–28
67. Elfenbein HA, Ambady N (2002) On the universality and cultural specificity of emotion recognition: a meta-analysis. Psychol Bull 128(2):203–35
68. Scherer KR, Clark-Polner E, Mortillaro M (2011) In the eye of the beholder? Universality and cultural specificity in the expression and perception of emotion. Int J Psychol 46(6):401–435
69. Li D, Rau PLP, Li Y (2010) A cross-cultural study: effect of robot appearance and task. Int J Soc Robot 2(2):175–186
70. Ekman P (1992) An argument for basic emotions. Cognit Emot 6(3–4):169–200