On Designing Expressive Robot Behavior: The Effect of Affective Cues on Interaction

Amir Aly, Adriana Tapus

Noname manuscript No.
(will be inserted by the editor)

On Designing Expressive Robot Behavior: The Effect of Affective Cues on Interaction

Amir Aly · Adriana Tapus

Received: date / Accepted: date

Abstract Creating a convincing affective robot behavior is a challenging task. In this paper, we are trying to coordinate between different modalities of communication: speech, facial expressions, and gestures to make the robot interact with human users in an expressive manner. The proposed system employs videos to induce target emotions in the participants so as to start interactive discussions between each participant and the robot around the content of each video. During each experiment of interaction, the expressive ALICE robot generates an adapted multimodal behavior to the affective content of the video, and the participant evaluates its characteristics at the end of the experiment. This study discusses the multimodality of the robot behavior and its positive effect on the clarity of the emotional content of interaction. Moreover, it provides personality and gender-based evaluations of the emotional expressivity of the generated behavior so as to investigate the way it was perceived by the introverted-extroverted and male-female participants within a human-robot interaction context.

1 INTRODUCTION

Robots are moving into human social spaces and collaborating in different tasks. An intelligent social robot is required to adapt the affective content of its generated behavior to the context of interaction and to the profile of the user in order to increase the credibility and appropriateness of its interactive intents. Speech, facial expressions, and gestures can express synchronized affective information that can enhance

Amir Aly
Ritsumeikan University
Japan
E-mail: amir.aly@ieee.org

Adriana Tapus
ENSTA ParisTech
France
E-mail: adriana.tapus@ensta-paristech.fr

behavior expressivity [18]. Gestures and facial expressions play an important role in explaining speech particularly in case of any speech signal deterioration [28].

Different studies in the literature of Human-Robot Interaction (HRI) and Human-Computer Interaction (HCI) discussed synthesizing affective speech [40, 58] and facial expressions [13, 65] in addition to gesture generation [20, 61]. Besides, other studies investigated the effect of multimodal information of speech and facial expressions on emotion recognition (compared to unimodal information) [17]. However, to our knowledge, these studies, among others, have not proposed a general framework to bridge between affective speech[1] (Section 3.1) on one side and both adaptive gestures [3, 4] and facial expressions (Section 3.2) on the other side, as illustrated in our current study. The proposed framework allows for an explicit control on prosody parameters so as to better express emotion. In addition, it considers the relationship between emotion and gestures, which allows for adapting the generated robot gestural behavior to the characteristics of the synthesized affective speech[2] according to the proposed context of interaction in this study. The illustrated system architecture in Section (3) guarantees a direct human-robot interaction context[3], which allows for generating and evaluating affective speech, adaptive gestures, and facial expressions so as to address the effect of the robot behavioral multimodality on interaction with a wide scope (Section 5.1). Additionally, we discuss another evaluation for the generated affective behavior of the robot based on the behavioral determinant factors of the participants: personality extraversion [33] and gender (Sections 5.2 and 5.3).

The important role that affective speech, gestures, and facial expressions could play in enhancing the robot behavior expressivity during social interaction is investigated through three experimental hypotheses of interaction between the participants and ALICE robot, where the robot behaviors with combined - at least two modalities of - speech, gestures, and/or facial expressions are compared to those with less affective cues[4] (Section 4.2). During the experiments, each participant watches a set of videos that aims at eliciting specific target emotions upon which interactive discussions with the robot start, where the participant evaluates the characteristics of the generated robot behavior (Section 5.1). Moreover, we report personality and gender-based evaluations for the robot behavior to find out any differences in the way it was perceived by the introverted-extroverted and male-female participants within a human-robot in-

[1] Mary-TTS, an open-source multilingual text-to-speech engine, is used to synthesize affective speech in the experiments.

[2] We generated adapted gestures [5, 81] to the synthesized affective speech instead of using human speech directly because *not all the participants are able to show an affective content in speech when describing a scene*, unless they are describing a personal experience they have been through (this describes the difference between *emotion perception* and *emotion experience* as explained in Schreuder et al. [74]), which is not the case in this study.

[3] Unlike the case if the participants were evaluating offline videos for the robot doing different behaviors without any interaction, which is out of interest in this study. Considering that we need to generate and model affective behavior on the robot, *we decided to create a context of affective interaction*. Consequently, we used videos with affective content from the database of Hewig et al. [36] whose content is centered around emotion elicitation as a base for interaction between each participant and the robot.

[4] For example, the robot behavior that employs combined speech, facial expressions, and gestures is compared to the robot behaviors expressed through speech only, speech and facial expressions, and speech and gestures so as to examine their effects on interaction.

teraction context so as to bridge between affective perception of the robot behavior and human profile (Sections 5.2 and 5.3). Last but not least, we discuss the findings of this study and propose research directions for future work (Section 6).

2 RELATED WORK

The correlation between emotion and speech has been extensively investigated in the related literature [26]. Speech prosody can reflect human emotion through variations in the basic features, like pitch, volume, and intensity [2, 59]. The variations in the characteristics of voice prosody that can influence the conveyed affective meaning of speech in case of different emotions, such as anger, disgust, fear, pleasure and sadness, were studied in Sauter et al. [72]. Emotion perception and the needed time for emotion recognition using prosodic features were discussed in Pell and Kotz [62].

The literature reveals different approaches towards synthesizing speech so as to improve both Human-Robot Interaction (HRI) and Human-Computer Interaction (HCI). Murray and Arnott [58] discussed a primary initiative to synthesize affective speech using a rule-based formant synthesis technique but the quality was low. Edgington [27] presented a concatenation based-technique that attained a little success in emotion expression. This last approach was further developed so that it employed the unit selection technique that avoids interference with the recorded voice to obtain a better quality of speech, and it reported some success in expressing anger, happiness, and sadness [40]. Similarly, deep learning approaches for speech synthesis have attracted attention over the last decade [31, 66, 92]; however, these approaches focused mainly on neutral speech synthesis. Moreover, end-to-end models (e.g., Tacotron model [90]) have been recently used in affective speech synthesis [51, 86]. However, these systems imitate a generic style of speaking in a few predefined emotions with a limited ability to control the affective expressivity of speech, which deprives them of flexibility and ease of use in our study considering the required large amount of data for training them. Generally, the previously discussed techniques, among others, do not have explicit control on the parameters of speech prosody to better express emotion. Therefore, in this work, we use the well-known pre-trained text-to-speech engine, Mary-TTS [75], to generate affective robot behavior expressed through speech (beside other modalities of communication, such as facial expressions and/or head-arm gestures) during interaction.

The basic definition of gesture was given by Kendon [45] and McNeill [56]. They defined a gesture as a synchronized body movement with speech, which is related in a parallel or complementary way to the meaning of an utterance. Ekman and Friesen [29] proposed a primary categorization of gestures: (1) affect displays (e.g., facial expressions), (2) adaptors (e.g., scratching), (3) regulators (e.g., using arm-hand movements to control turn-taking within a conversation), (4) illustrators (e.g., pointing), and (5) emblems (e.g., waving). This categorization was further adapted by Kendon [46] - due to neglecting language while it is a fundamental interactive phenomenon - who proposed a new gesture categorization: (1) signs (i.e., sign language), (2) pantomime (i.e., sequence of gestures with a narrative structure), (3) emblems, and (4) gesticulation. McNeill [56] named the continuum of Kendon's gesture categorization as *'Kendon's Continuum'* in his honor, and proposed another widely cited gesture

typology of four categories, which could be considered as gesticulations (according to Kendon's classification): (1) metaphorics (i.e., gestures referring to abstract ideas), (2) beats (e.g., rhythmic finger movements), (3) iconics (i.e., gestures with a close semantic correlation with speech that refer to images of specific entities), and (4) deictics (e.g., pointing). These categories represent the evolution of the described images and ideas in a speaker's mind.

The related literature in Human-Computer Interaction (HCI) and Human-Robot Interaction (HRI) shows active research towards generating iconic and metaphoric gestures that constitute a major part of the human nonverbal behavior during interaction [56]. Pelachaud [61] introduced the rule-based 3D agent GRETA that can generate a multimodal synchronized behavior using an input text. It can generate gestures of different categories regardless of the context and domain of interaction, contrarily to other 3D conversational agents (e.g., MAX agent) [48]. Cassell et al. [20] introduced a rule-based gesture generator; BEAT toolkit that can produce an animation script for both virtual agents (e.g., the agent REA) [19] and robots [7] from an input text. This toolkit can synthesize gestures of different categories such as iconic gestures, except for metaphoric gestures. Le et al. [50] proposed a rule-based framework for generating synchronized multimodal behaviors using the agent GRETA and robots. Generally, the majority of the rule-based gesture generation approaches do not consider the effect of emotion on body language, which could introduce a difficulty when adapting the generated robot behavior to human emotion detected through speech prosody [57] and gesture characteristics. Similarly, several deep learning approaches focused, increasingly, on gesture synthesis over the last years. Chiu et al. [22] proposed a data-driven framework for predicting gestures from speech; however, the model uses only predefined categories of annotated gesture data, which limits the shape of the produced gestures to those used in training with their language dependencies. Moreover, the model outputs gesture category labels rather than motion curves; therefore, it can not be used directly with 3D agents and robots. Hasegawa et al. [34] discussed a data-driven model for metaphoric gesture motion synthesis for a stick figure based on a speech input in Japanese; however, the generated gestures were rated relatively lower than the original gestures in semantic consistency. This model was further improved through motion representation learning to ameliorate gesture motion synthesis [49] but using the same language. Yoon et al. [91] introduced a data-driven end-to-end robot model for generating different categories of gestures (including iconic and metaphoric gestures) based on an input text and not a direct speech, which is similar to the rule-based gesture generators explained earlier. Besides, this model requires a very large amount of data for training. Therefore, in this paper, we present a complementary human-robot interaction study to our work [4] that discussed a framework for generating arm and head gestures adapted to speech prosody that correlates with emotion. These gestures are modeled on the robot in parallel with affective speech and/or facial expressions to examine the effect of the robot behavioral multimodality on interaction with human users.

The correlation between speech and facial expressions has been extensively investigated in the literature. Kalra et al. [41] showed that speech prosody and the movement of face muscles can change in a synchronous manner to express different emotions.

The unimodal perception of human emotion through audio or visual information was discussed in Silva et al. [79]. Additionally, Busso et al. [17] discussed the complementarity and combination of both modalities that can increase the perception of human emotion. Karras et al. [43] presented a Convolutional Neural Network (CNN) model that can synthesize 3D facial animation from speech - in different languages - expressing emotion. Other deep learning approaches have been discussed in Taylor et al. [83] and Vougioukas et al. [88] for facial animation synthesis from speech. These approaches, among others, are mostly limited to animating face models without focusing on generating facial expressions in different affective states.

In robotics and computer-based applications, modeling and synthesis of facial expressions have attracted much attention over the last decades. Platt and Badler [65] discussed a 3D face model that controls the responsible muscular actions for facial expressions following the Facial Action Coding System (FACS). Spencer-Smith et al. [80] presented a realistic 3D face model that can create different stimuli with 16 FACS units. Modeling credible facial expressions on robots was a rich topic of research in the last years due to their mechanical constraints compared to virtual agents that have a higher flexibility in creating facial expressions. Breazeal [15] presented the robot-head Kismet that employs eyes, mouth, and ears to model different emotions expressing sadness, surprise, happiness, disgust, and anger. Breemen et al. [16] introduced the robot iCat that can express fear, anger, sadness, and happiness. Beira et al. [13] developed the iCub robot that can model different emotions using gestures and facial expressions, such as happiness, anger, surprise, and sadness. Lutkebohle et al. [52] presented the robot-head Flobi that can express different emotions, such as fear, anger, surprise, sadness, and happiness. Hoffman et al. [37] developed the conversation companion Kip1, which can reflect emotion using a few degrees of freedom, like expressing fear through a shivering motion. Similarly, designing facial expressions on android robots has been a subject of extensive research to investigate the way to create convincing facial expressions considering the rules of human emotion expression [64]. Vlachos and Schärfe [87] investigated designing facial expressions on an android robot, where the findings showed the incapability of the robot to reproduce the ‘fear’ and ‘disgust’ emotions due to mechanical limitations in the face. These previous approaches for modeling facial expressions on 3D agents and robots, among others, show serious efforts towards creating expressive facial behaviors with specific emotions, and they report in the same time some limitations when modeling emotions with a wide scope. This indicates the importance of the robot behavioral multimodality, where each behavior modality enhances the other modalities so as to improve the clarity of the robot behavior during interaction.

The robot behavioral multimodality refers to coordinating and combining different modalities of communication in the robot (agent) behavior, which has been a challenging research topic over the last years [38, 84]. In facial expressions and gestures coordination, among others, Clavel et al. [23] discussed the positive effect of facial and bodily expressions on the affective expressivity of a virtual character (and consequently emotion recognition), and Costa et al. [25] proved that gestures can effectively help in recognizing the facial expressions of a robot. In speech and gestures coordination, among others, Salem et al. [71] discussed the positive effect of gestures

and speech multimodality on the evaluation of the robot behavior. In speech, gestures, and facial expressions coordination, among others, Castellano et al. [21] and Schirmer and Adolphs [73] reported the positive effect of multimodal information on emotion recognition compared to less-modal information. The related literature on the affective expressivity of the robot behavior has largely focused on unimodal (and bimodal) behaviors [38] considering the difficulty to generate a synchronized multimodal behavior, compared to virtual agents, with reasonably expressive speech, facial expressions, and gestures. This is due to the limited facial expressivity of robots that restricts generating a wide range of credible facial expressions, mechanical limitations that restrict generating gestures smoothly, and inability to synthesize affective speech for a wide range of emotions. In this work, we try to take a step forward towards creating a multimodal framework for generating affective robot behavior with more than two combined modalities of communication. Besides, we propose designs for modeling affective speech and facial expressions, in addition to gestures[5], which can inspire other researchers in social robotics with solutions when examining hard-to-model emotions. Furthermore, we discuss the participants' evaluations of the generated robot behavior considering their gender and personality, which is useful for future studies in human-robot interaction.

In this paper, we use the expressive ALICE robot for the purpose of modeling and evaluating a multimodal robot behavior expressed through combined, at least two modalities of, speech, facial expressions, and/or head-arm gestures compared to the robot behaviors with less combined affective cues. The paper is organized as follows: Section (3) discusses the system architecture, Section (4) illustrates the experimental hypotheses, design, and scenario of interaction, Sections (5 and 6) provide a description of the experimental results and a discussion of the outcome of the study, and finally, Section (7) concludes the paper.

3 SYSTEM ARCHITECTURE

This study presents a series of interaction experiments between humans and a robot, where the generated gestures and facial expressions of the robot depend on the synthesized affective speech (Figure 1) so as to create a multimodal affective robot behavior. The proposed framework is coordinated through the following subsystems:

1. **Speech Recognition**, which is the HTML5 multilingual Google API.
2. **Emotion Detection**, where predefined emotion-referring keywords are detected in the recognized speech of the participant, which correspond to his/her opinion about the projected video during each interaction experiment so as to label the emotional content of each video[6].

[5]In the proposed framework (Figure 1), facial expressions and gestures are generated adaptively to speech.

[6]The robot asks the participant to express his/her opinion about the content of a projected video[7]. Afterwards, it detects and segments predefined keywords, in a dictionary, from the comment of the participant, such as "This is ***disgusting***!" or "This video is expressing ***sadness***!". This helps in detecting the video's emotional content (from the participant's point of view) in order to trigger an adaptive robot behavior.

[7]We used a video database for emotion induction in the participants [36]. More details are available in Section (4.1).

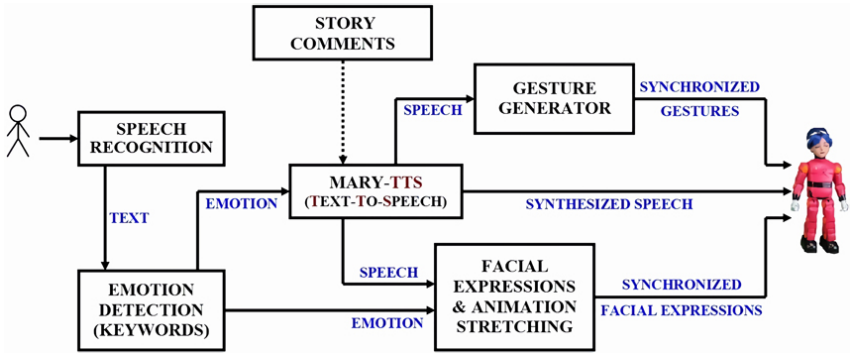


Fig. 1: Overview of the system architecture

3. **Mary-TTS Engine**, which converts the *story texts*[8] with the detected emotion labels of the employed videos to affective speech (Section 3.1).
4. **Body Gesture Generator**, which uses the generated speech by Mary-TTS engine to generate synchronized head-arm gestures[9] [4].
5. **Facial Expressions Modeling**, where facial expressions are modeled on the robot face in synchrony with the synthesized speech (Section 3.2).
6. **ALICE Robot**, which is the test bed platform in the conducted experiments with the participants (Section 4).

In the following sections of the paper, we illustrate the subsystems of the proposed framework and describe the experimental setup in detail.

3.1 Affective Speech Synthesis

The text-to-speech Mary-TTS engine is used for adding prosody and accent cues to a predefined text, which summarizes the storyline of a video under discussion [75]. This engine could help in making the robot able to engage in conversation with each participant using adaptive affective speech to the displayed story in the video. Mary-TTS engine uses a high-level markup language (SSML: Speech Synthesis Markup Language) to define the vocal pattern of the synthesized speech [82] as it provides different efficient features such as adding periods of silence between words in addition to providing an easy control on speech characteristics (i.e., pitch contour and baseline, and speech rate) (Figure 2). This could make it a helpful tool for the vocal design of the target emotions described in this study. It should be recalled that Mary-TTS engine is not yet prepared for synthesizing emotional speech in English

[8] **Story Comments**: are the predefined comments of the robot on the employed videos in the experiments. These story texts help in creating an interaction context between the participant and the robot associated with an adapted robot behavior - combining at least two modalities of emotional speech, facial expressions, and/or gestures - to the affective content of each video.

[9] This provides an implicit validation for the expressivity of the synthesized speech in which the more natural it is, the more natural will be the generated gestures (to be evaluated by the participants).

```
<?xml version="1.0" encoding="UTF-8"?>
- <speak xml:lang="en-US" xsi:schemaLocation="http://www.w3.org/2001/10
  /synthesis http://www.w3.org/TR/speech-synthesis/synthesis.xsd"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xmlns="http://www.w3.org/2001/10/synthesis" version="1.0">
   - <p>
      - <prosody rate="-30%" pitch="-4st" contour="(0%,+0st)(100%,-0st)">
            The video's content is so bad
            <break time="0.3s"/>
            innocent people have been attacked by policemen
            <break time="0.3s"/>
            who killed and injured a lot.
        </prosody>
    </p>
</speak>
```

Fig. 2: SSML specification of the 'sadness' emotion

in a human-like manner (same as other TTS engines); however, to our knowledge, Mary-TTS engine provides better vocal design capabilities and a higher flexibility than the other available engines. This makes the proposed vocal design in this work as an approximate step towards communicating the meaning of each expressed emotion during interaction. Thus, the robot behavioral multimodality is important for emphasizing the meaning of the expressed behavior, where each modality enhances the expressiveness of the other modalities.

Table (1) illustrates the proposed vocal patterns of the target emotions in which pitch contours are characterized by sets of parameters inside parentheses[10]. Speech rates of the target emotions vary between the rates of the 'sadness' emotion (lowest rate) and the 'anger' emotion (highest rate). The inter and intra-sentence break times were imposed experimentally on the proposed vocal design in order to enhance the affective expressivity of speech. The indicated inter-sentence break time with each emotion represents the silence periods that separate sentences at which both the lips and jaw of the robot make particular expressions to clarify the expressed emotion (Section 3.2). Besides, the intra-sentence break time indicates the silence periods of short duration within a sentence, which are necessary to clarify the expressivity of the 'sadness' and 'fear' emotions. The experimental parameters shown in Table (1) are an example of the prosody patterns of parts of the texts converted to speech for each emotion. The vocal patterns of the remaining parts of the texts differ slightly with respect to the indicated parameters in Table (1) so as to further clarify tonal variation over the text. Some emotions required using interjections (with tonal stress) in order to enhance their expressivity, like 'Ugh' and 'Yuck' for the 'disgust' emotion, and 'Oh my God' for the 'fear' emotion.

3.2 Facial Expressivity

The proposed design of facial expressions for the target emotions is grounded on the well-known coding system of facial actions (FACS) [30]. This design is clearly explained in Table (2), which shows the corresponding joints to each emotion in the face of the robot and the designed gestures to clarify the meaning of facial expres-

[10]The first parameter in each set followed by *"%"* denotes a percentage of the text duration, while the second parameter followed by *"st"* denotes the associated variation in baseline pitch in semitone.

Table 1: The design of the vocal pattern and contour behavior of each target emotion

Emotion	Baseline Pitch	Pitch Contour	Speech Rate	Contour Features			Break Time
				Start	Behavior	End	
Sadness	-4st	(0%,+0st)(100%,-0st)	-30%	Negative	Constant	Negative	Inter/Intra-Sentence
Disgust	+4st	(0%,-5st)(40%,-9st)(75%,-12st)(100%,-12st)	+8%	Negative	Exponential	Negative	Inter-Sentence
Happiness	+2st	(0%,+8st)(30%,+16st)(50%,+14st)(100%,+11st)	+7%	Positive	Parabola	Positive	Inter-Sentence
Anger	+5st	(0%,-18st)(50%,-14st)(75%,-10st)(100%,-14st)	+12%	Negative	Parabola	Negative	Inter-Sentence
Fear	+6st	(0%,+2st)(50%,+5st)(75%,+8st)(100%,+5st)	+7%	Positive	Parabola	Positive	Inter/Intra-Sentence

sions. The corresponding FACS units to emotions, in bold font, represent the most observed prototypical units between subjects [76], whereas the other units are observed at lower percentages. The underlined action units are the units with corresponding relative joints in the face of the robot.

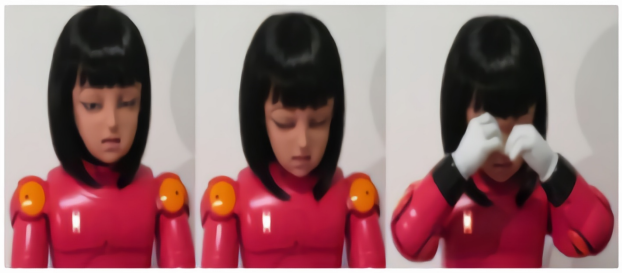
Table 2: The design of facial expressions, modeled on the robot, for each target emotion

Emotion	FACS Coding	Robot Face Joints	Additional Body Gestures
Sadness	**Brow Lowerer** + **Lip Corner Depressor** + Inner Brow Raiser + Cheek Raiser + Nasolabial Deepener + Chin Raiser	Left Smile + Right Smile + Brows	*Covering-Eyes Hand* + *Bowing Head* + Narrowing Eyes + Eyes Blinking + Closing Jaw
Disgust	Lip Pressor + Brow Lowerer + **Nose Wrinkler** + **Upper Lip Raiser** + **Chin Raiser**	Jaw + Brows	*Neck Rotation* + *Raising Front-Bent Arms* + Narrowing Eyes
Happiness	**Lip Corner Puller** + **Lips Part** + Cheek Raiser	Left Smile + Right Smile + Jaw	Eyes Blinking
Anger	**Brow Lowerer** + **Lid Tightener** + **Lip Pressor** + Lip Tightener + Upper Lip Raiser + Chin Raiser + Nasolabial Deepener	Jaw + Brows + Eyelids	*Down Head-Shaking* + Short Mouth-Opening
Fear	**Inner Brow Raiser** + **Brow Lowerer** + **Lip Stretcher** + **Lips Part** + Outer Brow Raiser + Upper Lid Raiser + Jaw Drop	*Left Smile* + *Right Smile* + Jaw + Brows + Eyelids	*Mouth-Guard Hand*

The complexity behind modeling emotion on the face of the robot lies in the absence of equivalent joints to specific FACS descriptors (e.g., cheek raiser and nose wrinkler). Therefore, and inspired by the experimental designs of McColl and Nejat [55] and Wallbott [89][11], we imposed some additional body gestures experimentally in order to reduce the negative effect of the absent joints on affective expressivity. These additional gestures do not include neither head gestures nor arm-hand gestures, which are generated by the gesture generator [4] (except for the italic-font gestures indicated in Table (2), which are required to enhance the affective expressivity of the robot)[12]. For example, the combination of the additional gestures *neck rotation* and *raising front-bent arms* is helpful for better expressing the 'disgust' emotion (Figure 3), which can give the participant the feeling that the robot does not like the interaction context. In a similar way, the emotions of 'sadness', 'fear', and 'anger' are assigned the gestures of *bowing head and covering-eyes with hand*, *mouth-guard with hand*, and *down head-shaking*, respectively, to emphasize their affective expressivity (Figure 3). The main role of the additional *right smile* and *left smile* face joints

[11] These studies discuss the characteristics of body behavior in different emotions employing arm gestures. McColl and Nejat [55] used the gesture *hanging arms* to express the sadness emotion using the robot Brian-2, while Wallbott [89] used the gesture *crossed in front of chest* to describe the disgust emotion. The final implementation of these gestures on ALICE robot was made according to the mechanical limitations of the robot arms.

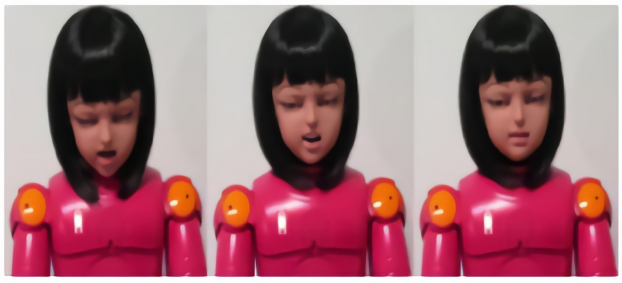
[12] The metaphoric gesture generator [4] synthesizes the most appropriate head-arm gestures based on its own learning algorithm. Consequently, it is possible that the predefined additional gestures (in italic font, Table 2) might not be generated during the interaction. Thus, we added them, experimentally, at particular moments of speech with a higher priority than the synthesized gestures by the generator.



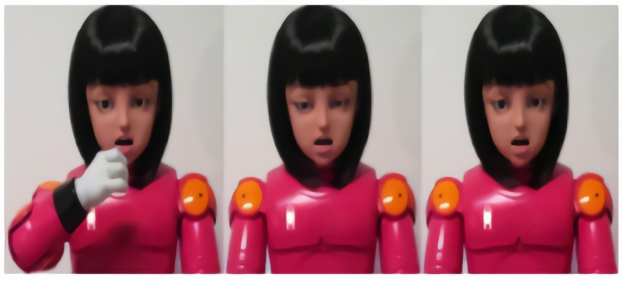
(a) Sadness

(b) Disgust

(c) Happiness



(d) Anger



(e) Fear

Fig. 3: Synthesized facial expressions by ALICE robot

```
<?xml version="1.0"?>
- <Animation>
  - <Version type="Animation">
      <Name>New Animation</Name>
      <Number>1.0</Number>
    </Version>
  - <Channels>
    - <Channel name="Eyelids" id="301">
      - <MotionPaths>
        - <MotionPath name="path">
          - <Version type="Interpolation">
              <Name>C-Spline Interpolation</Name>
              <Number>1.0</Number>
            </Version>
          - <ControlPoints>
            - <ControlPoint>
                <Time>195.0</Time>
                <Position>0.4991511035653651</Position>
              </ControlPoint>
            - <ControlPoint>
                <Time>1270.0</Time>
                <Position>0.9252971137521222</Position>
              </ControlPoint>
            - <ControlPoint>
                <Time>2395.0</Time>
                <Position>0.12224108658743638</Position>
              </ControlPoint>
            </ControlPoints>
          </MotionPath>
        </MotionPaths>
      </Channel>
    </Channels>
</Animation>
```

Fig. 4: Eyelids animation script

of the 'fear' emotion is to depress the corners of the open mouth so as to enhance its affective expressivity; however, both joints do not have equivalent FACS descriptors (Table 2). Generally, modeling persuasive facial expressions on a robot is not a trivial task because of the mechanical limitations of its joints (unlike the case with 3D agents). Therefore, the robot behavioral multimodality can play an important role in enhancing its affective expressivity during interaction, where each behavior modality can clarify the other modalities.

Figure (4) demonstrates the eyelids animation script where three points of the motion path are described through position and time. In order to achieve a temporal alignment between eyelids animation and speech, if the synthesized speech duration is longer or shorter than the eyelids animation duration, the model determines the corresponding new time instants to animation points based on speech duration, animation duration, and the previous time instants of animation points. The segmentation of human speech is achieved through an embedded voice activity detection algorithm in the speech recognition system, which can efficiently label speech and silence segments. In case the silence period represents an inter-sentence break time that was discussed in Section (3.1), both of the robot jaw and lips perform specific animations (e.g., pulling the corners of the lips to express happiness) which could enhance the meaning of the expressed emotion (Figure 3). This is due to the robot mechanical constraints that prevent the synchronization between lips motion and speech while performing an animation with both the jaw and lips at the same time. Meanwhile, if the silence period corresponds to an intra-sentence break time, the jaw of the robot opens to express fear and closes to express sadness during the silence period (Section 3.1).

Table 3: The target emotions and their corresponding feature films. The main videos were extracted from the bold-font films. Meanwhile, the other films represent the standby videos.

Target Emotion	Feature Film
Sadness	**The Champ -** An Officer and a Gentleman
Disgust	**Pink Flamingos -** Maria's Lovers
Happiness	**On Golden Pond -** An Officer and a Gentleman
Anger	**My Bodyguard -** Cry Freedom
Fear	**Halloween -** Silence of the Lambs
Neutral	**Crimes and Misdemeanors -** All the President's Men

4 EXPERIMENTAL SETUP

In this section, we discuss the employed database for emotion induction in the participants. In addition, we present the experimental hypotheses, design, and scenario of interaction between the participant and ALICE robot developed by RoboKind[13].

4.1 Database

The employed database contains 20 silent videos excerpted from feature films (with duration varying from 29 to 236 seconds) for inducing 6 target emotions in the participants: neutral, disgust, anger, happiness, fear, and sadness[14]. Hewig et al. [36] discussed and validated the efficiency of the database in eliciting emotions in humans. Consequently, in this paper, we will not focus on measuring the level of emotion induction in the participants[15]. During the experiments, we used 12 expressive videos from the database to elicit the target emotions. This means that six main videos were used during the experiments, and six standby videos (i.e., one standby video per emotion) were used automatically in case any of the main videos failed to elicit the corresponding target emotion (Table 3).

4.2 Hypotheses

Human emotion experience is generally characterized by different cognitive constructs, such as (1) emotion clarity (i.e., the clear and definite representation of emotion) [24], (2) emotion differentiation, which is the ability to accurately identify and represent emotion into discrete categories (e.g., sadness, disgust, and happiness). This is conceptually correlating with emotion clarity, where each construct could enhance the other one [14], (3) emotional complexity (i.e., the broad range of emotion experiences associated with a tendency to accurately differentiate between emotion categories) [42], and (4) emotional awareness (i.e., the knowledge complexity of emotion,

[13]The humanoid ALICE-R50 robot has an expressive face and a total of 36 degrees of freedom in the whole body. The robot has two cameras and a sensor set in order to perceive its surrounding environment. The robot face with synthetic skin can efficiently make a variety of facial expressions with high credibility (Section 3.2).

[14]The **surprise emotion** is not considered in this study because it is not included in the video database of Hewig et al. [36], which we used for emotion induction.

[15]We correlate between emotion induction and recognition using videos so that an induced emotion could be correctly recognized by the participant.

which represents the ability to be aware of emotion) [54]. Each of these constructs is measured through calculated indices from subjects' self-reports [44].

In this research study, the main objective is to generate a well-perceived multimodal robot behavior so as to enhance the interaction with a human user. Consequently, the clarity and differentiation constructs of emotion would be directly addressed through investigating the ability of the participants to recognize the affective content of the generated robot behavior[16]. Besides, the participants would evaluate the effect of the robot behavioral multimodality on interaction. The subjective evaluation of the generated multimodal robot behavior investigates basically the *clarity*[17]/ *expressivity*[18], and the *recognizability* (i.e., *emotion differentiation*) of the affective robot behavior in addition to the synchronization between the behavior modalities, etc. The examined hypotheses in this study are:

- **H1:** The combination of facial expressions, speech, and arm and head gestures will increase the clarity of the affective content of the robot behavior to the participant compared to the experimental conditions with less combined affective cues (i.e., less combined modalities of communication).
- **H2:** Facial expressions will enhance the recognizability, and expressivity, of the robot emotion by the participant compared to the experimental conditions without facial expressions.
- **H3:** The characteristics of the arm and head gestures of the robot (e.g., acceleration) will enhance the expressivity of the robot behavior so as to help the participant in recognizing and distinguishing between emotions compared to the experimental conditions without arm and head gestures.

The effect of emotional speech on interaction was not examined through an independent hypothesis because this requires whether:

- Comparing the robot behavior that employs affective speech to the robot behavior that does not employ affective speech (i.e., using neutral or monotone speech). However, the proposed system in this study uses the synthesized speech as a basis for generating synchronized gestures with facial expressions (Figure 1). Therefore, synthesizing monotone speech will lead to associated facial expressions and gestures with *different characteristics* than those of the facial expressions and gestures generated using affective speech. Consequently, it is not possible to compare between the robot behaviors in similar experimental conditions (e.g., the robot behavior expressed through speech and gestures in the case of affective speech and

[16]This is based on their previous experiences with the target emotions, which are common and basic emotions that each person whether experiences internally or perceives through speech, facial expressions, and gestures of others in the environment.

[17]The maximum possible level of emotional expressivity achieved through combining speech, facial expressions, and head-arm gestures together (Figure 6), which is concordant with the definition of **emotion clarity** discussed earlier in the same section [24].

[18]A lower level of emotional expressivity achieved through less affective cues than in the clarity level (Figure 6).

the same behavior in the case of monotone speech as gestures in both cases will be different).

– Comparing the robot behavior that employs affective speech to the robot behavior that does not employ speech at all. This condition does not match the context of the *non-mute* human-robot interaction[19].

Consequently, these two cases are excluded from our experimental design. Instead, the important role of speech in enhancing the affective content of interaction would be measured directly through analyzing the post-experiment questionnaires.

4.3 Experimental Design

The experimental design is based on the *between-subjects* design[20] through a human-robot interaction context in which the synthesized speech by Mary-TTS (text-to-speech) engine is used as an input to the gesture generator [4] so as to synthesize adapted gestures to the synthesized affective speech[21]. This constitutes an implicit validation for the expressivity of the synthesized speech using Mary-TTS engine in which the more natural (i.e., human-like) the synthesized speech is, the more natural will be the corresponding generated gestures (to be evaluated by the participants). Besides, generating adaptive gestures based on speech characteristics is concordant with the cognitive co-production process of synchronized speech and gestures that humans undergo [56]. The synthesized speech and gestures (in addition to facial expressions) are modeled on the robot and evaluated by the participants at the end of each conducted experiment. The proposed design includes the following robot behavior conditions:

– The robot produces a multimodal affective behavior expressed through facial expressions, speech, and arm and head gestures (i.e., condition C1-SFG).

– The robot produces a multimodal affective behavior expressed through facial expressions and speech (i.e., condition C2-SF).

– The robot produces a multimodal affective behavior expressed through arm and head gestures, and speech (i.e., condition C3-SG).

– The robot produces a unimodal affective behavior expressed through speech (i.e., condition C4-S).

In order to validate the first hypothesis, the experimental conditions C1-SFG, C2-SF, C3-SG, and C4-S were examined. While for the second hypothesis, the conditions

[19] This study is focusing on investigating the effect of the robot behavioral multimodality on interaction with typically developed individuals who use speech, facial expressions, and gestures for daily communication. Consequently, excluding speech from interaction will certainly hinder conveying messages (using only facial expressions and/or gestures) in a normal manner unless we use a conventionalized sign language in parallel, which is totally away from the scope of the current study.

[20] Each experimental condition is evaluated through a different group of participants.

[21] We used the synthesized affective speech by Mary-TTS engine to generate a robot gestural behavior instead of using human speech directly because not all the participants are able to show an affective content in speech when describing a scene, unless they are describing a personal experience they have been through (this describes the difference between *emotion perception* and *emotion experience* as explained in Schreuder et al. [74]), which is not the case in this study.

Fig. 5: Interaction experiments between the robot and two different participants

C2-SF and C4-S were examined, and for the third hypothesis, the conditions C3-SG and C4-S were examined. We excluded the condition of the robot producing a unimodal behavior expressed through facial expressions or arm and head gestures without using speech, and the condition of the robot producing arm and head gestures combined with facial expressions without using speech (Section 4.2). The condition C3-SG was excluded from validating the second hypothesis and the condition C2-SF was excluded from validating the third hypothesis because the facial expressions of the robot are associated with the additional body gestures detailed in Table (2). Consequently, separating between the conditions of facial expressions and gestures (i.e., conditions C2-SF and C3-SG) could guarantee differentiating between the accompanying gestures to the robot facial expressions and the basic head-arm gestures synthesized by the generator. This could lead to better evaluating the effect of facial expressions and gestures on interaction.

The literature reveals serious efforts to elicit emotion in humans under laboratory conditions. These emotion induction methods include: dyadic interaction tasks [70], affective imagery [47], music [69], and pictures and film clips [85]. In this study, the robot and the participant, in each condition, follow an expressive stimulus set of short videos through six experiments that mean to elicit six different target emotions (Figure 5) after a short preparation phase[22]. The scenario of interaction is described as follows:

- The robot invites the participant to watch some videos and discuss their storylines.
- The robot asks the participant to express his/her opinion about the content of the projected video. Afterwards, it detects and segments predefined emotion-referring

[22] **Pre-Experiment Preparation Phase**:The experimenter introduced the humanoid expressive ALICE robot to the participant and explained the task. Each participant signed an informed consent to be notified about different points such as nature of the study, duration of interaction[23], data privacy, statement of risks and benefits, right to get informed about results in addition to giving an authorization to get filmed. The participant was seated in front of the robot with a table in-between, and used a headset microphone to capture his/her own speech during interaction [6].

[23] Each experiment had a varying duration between 1 and 4 minutes, while the duration of answering each questionnaire was varying between 2 and 5 minutes.

Table 4: The recognition scores of the videos' affective contents in the different experimental conditions

Condition	Emotion Induction	
	Correct (after the 1st videos)	Correct (after the 2nd videos)
C1-SFG	100%	0%
C2-SF	100%	0%
C3-SG	98.9%	1.1%
C4-S	97.8%	2.2%

keyword(s) from the recognized comment of the participant, such as "This is *disgusting*!", "This video is expressing *sadness*!", etc. This helps in detecting the video's emotional content (from the participant's point of view) in order to trigger a corresponding adaptive robot behavior.

- After listening to the participant's comment on the video, the robot makes a comment accompanied by speech, facial expressions, and/or head-arm gestures on the content of the video.
- If the displayed video induces, in the participant, another emotion than the concerned target emotion so that the system detects keyword(s) that belong mainly to another category of emotion-referring keywords, the robot comments through a *neutral* behavior. Thereupon, the robot asks the participant to watch a different video so as to retry to induce the emotion that was failed to be elicited using the first video (Table 4).
- The experiment terminates for the examined target emotion. Thereupon, the participant evaluates the generated behavior of the robot through a 7-point Likert scale questionnaire. This evaluation focuses on the relevance of the robot behavior to the context of interaction in terms of its emotional content and expressivity, synchronization between the robot behavior modalities (i.e., speech, facial expressions, and/or gestures according to the examined experimental condition), etc [24]. Afterwards, a new experiment of interaction starts for examining a different, randomly selected, target emotion.
- After all the experiments terminate, the experimenter and the robot express gratitude to the participant for his/her time and cooperation.

Table (4) shows that the majority of the target emotions were correctly recognized by the participants after watching the first videos in the four experimental conditions, while the second videos were slightly required. This shows that the chosen videos from the employed *silent* video database[25] had convincing emotional contents [36].

[24] An example of a Likert scale question that evaluates the clarity of the robot behavior during the conducted experiments (1 $\longrightarrow$ lowest score, 7 $\longrightarrow$ highest score):

- How do you evaluate the affective expressivity of the generated robot behavior?

	1	2	3	4	5	6	7
Low (Not Clear Emotion)	☐	☐	☐	☐	☐	☐	☐

[25] This *silent* video database was created for serving brain asymmetry research to avoid affecting asymmetry measures with speech, sound, and music [36].

Afterwards, the participants were first asked through each post-experiment questionnaire to evaluate the characteristics of the generated robot behavior in terms of each modality of communication (i.e., speech, gestures, and facial expressions) independently, then they were asked to evaluate and recognize the affective content of the generated combined behavior. We argue that this supports separating between the emotional contents of the videos and the robot behaviors during evaluation - supported by the findings of Hermans et al. [35][26] - up to the level that allows for investigating the experimental conditions successfully[27].

5 EXPERIMENTAL RESULTS

A total of 60 participants were recruited in order to validate the different examined hypotheses in this study. The participants have been equally distributed over the experimental conditions (i.e., 6 females and 9 males for every condition). The participants were undergraduate and postgraduate students and employees at ENSTA-ParisTech (with ages varying from 20 to 57 years old, $M = 29.6$ and $SD = 9.4$). The participants had a technical background with an average of 66.7%, and a non-technical background with an average of 33.3%. Moreover, only 40% of the participants had previous interaction experience with robots, while 60% of them did not interact with robots beforehand. The effect of synthesizing adaptive robot behavior on interaction with the participants in addition to personality and gender-based evaluations of the emotional expressivity of the generated behavior are illustrated in the following points:

5.1 Effect of the Robot Behavioral Multimodality on Interaction

For the first hypothesis, a significant difference was found by ANOVA analysis in the clarity of the affective robot behavior expressed through a combination of speech, facial expressions, and head-arm gestures with respect to the robot behaviors, with less affective cues, expressed through speech, speech and facial expressions, and speech and head-arm gestures ($F[3,356] = 21.15$, $p < 0.001$) (Figure 6). Tukey's HSD comparisons indicated a significant difference in clarity between the robot behavior expressed through combined speech, facial expressions, and head-arm gestures (i.e.,

[26]Hermans et al. [35] argued that affective priming results from fast-acting cognitive processes whose **effects quickly dissipate after a short duration** of milliseconds.

[27]According to the study of Schreuder et al. [74], *emotion perception* results from the interpretation of the emotional qualities of the stimulus, while *emotion experience* is a state that results from the internal assessment of the percept. This means that a human might perceive a stimulus with emotional content (with/without) experiencing any internal emotions depending on the stimulus, the context, and the person. *Emotion elicitation* is the intermediate phase that links between emotion perception and emotion experience. The employed database in the experiments had been evaluated with emotional eliciting content as discussed in Hewig et al. [36]. However, as the process of emotion elicitation highly depends on the human and his/her previous emotional experience, it is very difficult to define the level of emotion elicitation in the recruited participants during the experiments so as to detect if it was sufficient to have any effect on the evaluation of the robot behavior. This needs another psycho-cognitive study and different experimental conditions to investigate. However, based on the findings of Hermans et al. [35], *we believe that evaluating the robot behavior was not influenced by the videos*. It might be important to notice that the participants evaluated the robot behavior freely regardless of the content of the videos so that when the robot behavior had a convincing affective content, it received a high evaluation, to the contrary of the case when it had a less convincing affective content, which supports our proposed experimental design.

Table 5: The scores of recognizing the target emotions, modeled on the robot, in different conditions

Condition	Emotion					
	Sadness	Disgust	Happiness	Anger	Fear	Neutral
C2-SF	100%	80%	93.3%	92.9%	100%	100%
C3-SG	100%	93.3%	93.3%	92.3%	100%	100%
C4-S	100%	93.3%	93.3%	80%	100%	100%

condition C1-SFG) on one side, and the robot behaviors expressed through speech (i.e., condition C4-S) ($p < 0.001$) (the lowest among the four conditions), speech and facial expressions (i.e., condition C2-SF) ($p < 0.001$), and speech and head-arm gestures (i.e., condition C3-SG) ($p < 0.001$) on the other side. Moreover, no significant difference was observed between the conditions C2-SF and C3-SG in the clarity of the robot behavior.

For the second hypothesis, the robot behavior expressed though facial expressions and speech was found by the participants to be more expressive and adapted to the context of interaction than the behavior expressed through speech ($F[1,178] = 18.63$, $p < 0.001$). Moreover, the participants considered that speech and facial expressions were synchronized with an average score of $M = 5.9$, $SD = 0.9$. Furthermore, they did not find any significant inconsistency in affective content between speech and facial expressions with an average score of $M = 1.8$, $SD = 1.2$. Over and above, they agreed that speech was less expressive than facial expressions with an average score of $M = 4.4$, $SD = 1.5$. Table (5) shows that facial expressions improved only the score of recognizing the emotion of 'anger' in the experimental condition C2-SF with reference to the condition C4-S, which is related to the limitations of Mary-TTS engine in designing a highly expressive vocal pattern for this particular emotion (Section 3.1), so that facial expressions enhanced the affective content of speech giving the participants the feeling that the robot was expressing the 'anger' emotion persuasively. To the contrary, the facial expressions of the robot had a negative effect on the score of recognizing the emotion of 'disgust' in the experimental condition C2-SF with reference to the condition C4-S, which is related to the limited affective facial expressivity for this particular emotion (Section 3.2).

For the third hypothesis, the affective content of the robot behavior expressed through both arm and head gestures and speech was considered to be more expressive and observable by the participants than that of the behavior expressed through speech ($F[1,178] = 17.16$, $p < 0.001$). Furthermore, the participants found that speech and gestures were synchronized with an average score of $M = 6.1$, $SD = 0.7$, and they agreed that the execution of gestures was fluid with an average score of $M = 5.35$, $SD = 1.03$. Over and above, the participants found that gestures were more expressive than speech with an average score of $M = 4.25$, $SD = 1.43$. The affective content of the arm and head gestures of the robot behavior was reasonably recognized by the participants (Table 5). The generated gestures ameliorated only the score of recognizing the emotion of 'anger' in the experimental condition C3-SG with reference to

the condition C4-S, which is related to gesture characteristics such as velocity and acceleration, that enhanced the robot expressivity for this emotion.

Figure (6) illustrates the variation in the affective expressivity of the robot behavior in the experimental conditions C1-SFG, C2-SF, C3-SG, and C4-S. The robot behavioral expressivity in each condition was investigated through a different group of 15 participants. The combination of different affective cues (i.e., speech, facial expressions, and head-arm gestures in the condition C1-SFG) provided clarity to the robot behavior with respect to the other conditions that employ less affective cues as argued in the first hypothesis[28]. Meanwhile, no significant difference was observed in the robot behavioral expressivity between the conditions C2-SF and C3-SG.

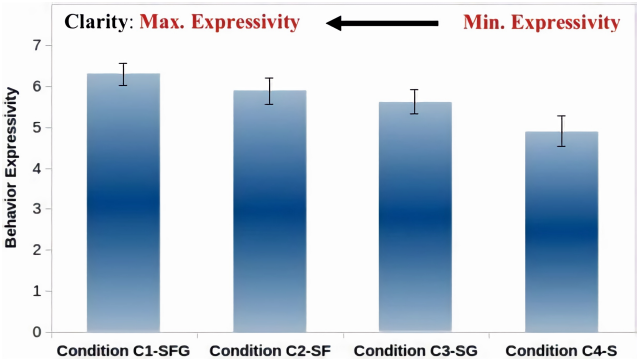


Fig. 6: Human perception of the emotional expressivity of the robot behavior in the four experimental conditions, where the *clarity* of behavior refers to the maximum level of expressivity it can show

A significant result was found by two-way ANOVA analysis in the perception of the affective robot behavior with clarity-expressivity of facial expressions (i.e., condition C2-SF) and emotion as independent variables ($F[2,168] = 4.47$, $p = 0.0359$). However, no significant result was found with clarity-expressivity of gestures (i.e., condition C3-SG) and emotion as independent variables. After running one-way ANOVA analysis on each emotion individually, the results showed that both the 'happiness' and 'disgust' emotions were found significantly more clear when being expressed through combined speech, facial expressions, and head-arm gestures (i.e., condition C1-SFG) ($F[1,28] = 3.36$, $p = 0.077$) than when being expressed though speech and facial expressions (i.e., condition C2-SF) ($F[1,28] = 6.133$, $p = 0.0196$). Meanwhile, no significant differences were found for the 'neutral', 'sadness', 'fear', and 'anger' emotions. Over and above, a statistically significant main effect was observed for the

[28] **Clarity** and **expressivity** have been previously defined in Section (4.2). An affective robot behavior could have some level of *expressivity*, but it could be *not* really clear to the participants in the same time. For example, the interpretation of a facial expression could be ambiguous and confused among different emotions (i.e., it is expressive, but not clear enough to be fully perceived), in this case speech or gestures could help in interpreting the actual emotion so as to enhance its *clarity*.

Table 6: The numbers of the introverted and extraverted participants in the four experimental conditions

Personality Dimension	Condition (15 Participants / Condition)			
	C1-SFG	C2-SF	C3-SG	C4-S
Introversion	6	6	8	7
Extraversion	9	9	7	8

experimental conditions ($F[3,335] = 12.738$, $p < 0.001$) and for the target emotions ($F[5,335] = 5.527$, $p < 0.001$).

5.2 Human Personality-Based Evaluation of the Affective Robot Behavior

Personality is a determinant factor in human social interaction, which has a long-term consistent effect on the generated multimodal human behavior. Reisenzein and Weber [67] defined personality as the coherent and collective pattern of emotion, cognition, behavior, and goals over time and space. Moreover, Revelle and Scherer [68] discussed the strong relationship between personality and emotion. Several research studies in neuroscience discussed the correlation between the neurobiological structure of personality extraversion and the activation in different brain regions involved in emotional responding (which implies perceiving the affective content of interaction) [39]. This potential correlation between personality extraversion and emotion perception would be investigated within a human-robot interaction context so as to study the effect of human personality on perceiving the emotional expressivity of the robot behavior.

5.2.1 Personality Extraversion-Based Evaluation of the Affective Robot Behavior

Table (6) indicates the numbers of the introverts and extraverts in each experimental condition, where the calculation of personality scores was based on the online Big5 personality model questionnaire [32][29] that each participant filled in at the beginning of the experiments. Figure (7) illustrates the effect of the human extraversion personality trait - in terms of the introversion and extraversion of personality - on the perception of the affective expressivity of the robot behavior. In the four experimental conditions, both the introverts and extraverts showed a similar tendency in evaluating the emotional expressivity of the robot behavior, where the perception of the extraverted participants for the robot behavior was, in general, higher than that of the introverted participants. The variance in evaluating the expressivity of the robot behavior by the introverted and extraverted participants was found statistically significant (through T-Test) in the different conditions: C1-SFG ($p < 0.02$), C2-SF ($p < 0.03$), C3-SG ($p < 0.03$), and C4-S ($p < 0.02$).

This evaluation difference between the introverted and extraverted participants is concordant with the findings of Shulman and Hemenover [77], Petrides et al. [63], and Atta et al. [12], who argued that emotional intelligence[30] is positively correlating with

[29] http://www.outofservice.com/bigfive/

[30] The ability to perceive others' emotions through analyzing the affective cues of their behaviors [53].

personality extraversion. Consequently, the extraverted participants are expected to have a *relatively* higher emotional intelligence than that of the introverted participants so that they gave higher ratings for the robot behavior in the four experimental conditions. The previous evaluation of the affective expressivity of the robot behavior matches the illustrated findings in Figure (6), where the evaluation of the robot behavior in the condition C1-SFG was higher than that in the other conditions.

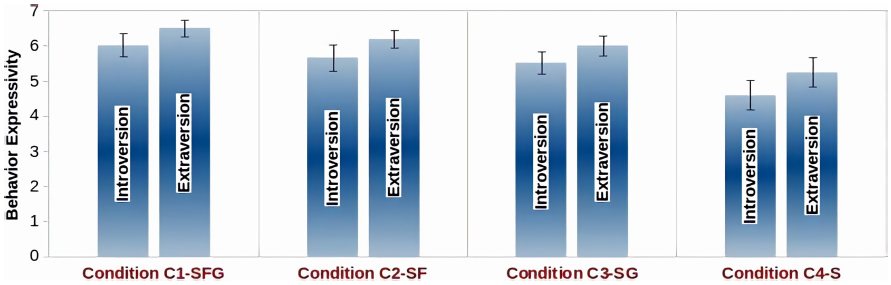


Fig. 7: Human personality-based evaluation (in terms of introversion and extraversion of personality) of the affective expressivity of the robot behavior

5.3 Gender-Based Evaluation of the Affective Robot Behavior

Both of the female and male participants have positively perceived the affective expressivity of the generated robot behavior in the four experimental conditions (Figure 8). The indicated ratings in the figure show that the perception of the male participants for the affective robot behavior in the four conditions was generally higher than the perception of the female participants. This relatively higher preference of the male participants over the female participants for the emotional expressivity of the female ALICE robot matches the findings of Siegel et al. [78] and Park et al. [60], where they found that the participants considered the opposite-sex robots to be more attractive and convincing during interaction.

The variance between the ratings of the male and female participants for the emotional expressivity of the robot behavior indicated in Figure (8) was found statistically significant (through T-Test) in the different conditions: C1-SFG ($p < 0.02$), C2-SF ($p < 0.03$), C3-SG ($p < 0.02$), and C4-S ($p < 0.001$). Furthermore, the male participants considered the generated multimodal robot behavior more adapted to the emotional content of the videos, and consequently the context of interaction, than the female participants ($p < 0.01$), which supports the hypothesis of the opposite-sex attraction of human users to robots.

The observable difference between the ratings of the male and female participants in the condition C4-S compared to those in the conditions C1-SFG, C2-SF, and C3-SG (Figure 8) could be related to the low affective expressivity of the robot behavior employing speech only in interaction with respect to those that employ speech combined with facial expressions and/or gestures (Figure 6). We argue that facial expressions and gestures enhanced the affective content of the robot behavior, which

slightly improved the perception of the female participants to the generated behavior in the conditions C1-SFG, C2-SF, and C3-SG while keeping the opposite-sex attraction hypothesis of human users to robots valid. These findings need; however, a larger number of male and female participants to have a clearer visualization for their perceptual differences of the robot behavior.

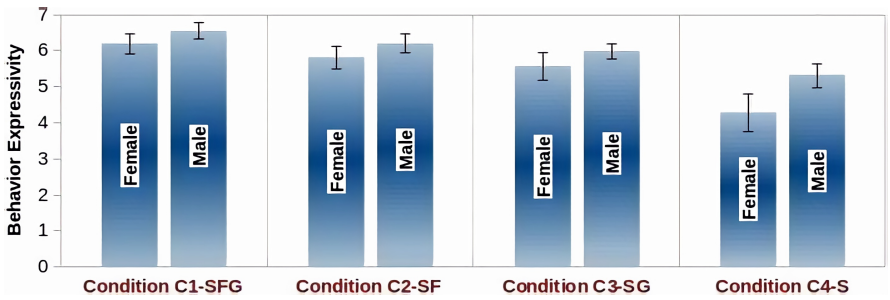


Fig. 8: Gender-based evaluation of the affective expressivity of the robot behavior

6 DISCUSSION

We propose an integrated system for generating affective robot behavior expressed through speech, gestures, and facial expressions within a human-robot interaction context. We investigate the multimodality of the generated robot behavior and its positive effect on interaction with the participants through three experimental hypotheses that compare between the robot behavior with combined, at least two modalities of, speech, gestures, and/or facial expressions and those with less affective cues. Moreover, we investigate any potential effect of human personality and gender on the way the robot behavior was perceived during interaction.

The proposed framework (Section 3) integrates different subsystems for affective speech synthesis, gesture generation based on speech prosody, and an expressive robot with highly credible facial expressions, which allows for studying the effect of the robot behavioral multimodality on interaction with a wide scope. The obtained results demonstrate the positive role that affective cues could play in enhancing the expressivity of the robot behavior so as to help the participants in perceiving its emotional content appropriately. These findings are clearly illustrated in Figure (6), where the robot behavior that combines speech, facial expressions, and gestures attained a higher level of expressivity (i.e., clarity level) than the other robot behaviors with less affective cues.

When searching in the related studies in the literature for concordant results with our findings on affect recognition using multimodal information, we found that the majority of them were unimodal (and bimodal) - based approaches employing, among others, gestures and facial expressions, speech and gestures, and speech and physiological signals [38, 93]. Meanwhile, there are a few studies that discussed emotion recognition with more that two modalities of information. Castellano et al. [21] used

speech, gestures, and facial expressions to recognize emotions, and reported that using multimodal data for affect recognition highly increased the scores with respect to the cases that use less modalities of data [73]. Generally, our proposed system shares the same concept of the positive effect of multimodality on emotion perception and recognition. However, it is designed to generate and *embody* a multimodal behavior - expressed through speech, gestures, and facial expressions - on ALICE robot so as to be positively perceived by the participants, which makes it a *different contribution than any other approach* in the related literature.

Over and above, the results report some differences in the perception of the introverted-extroverted and male-female participants for the affective robot behavior, where the perception of the extraverted and male participants for the robot behavior was generally higher than that of the introverted and female participants in the different conditions of behavior (Figures 7 and 8). While we tried to explain these findings in light of other similar findings in the related literature (Sections 5.2.1 and 5.3) so as to support our results, we believe that a larger number of introverted-extroverted and male-female participants is required in order to figure out their perceptual differences of the robot behavior more precisely. However, we argue that the current results could give useful insights into human perception of the affective robot behavior to the other interested researchers in the field of human-robot interaction.

7 CONCLUSION

This paper introduces a framework for generating an adapted multimodal robot behavior, expressed through speech, gestures, and/or facial expressions, to the context of interaction with human users. A set of videos that mean to induce target emotions in the participants is employed during the experiments upon which interactive discussions start with the robot around their affective contents. Each participant is only exposed to one of the four experimental conditions of multimodal - unimodal robot behaviors during the experiments. The system uses Mary-TTS engine to generate emotional speech; however, the proposed vocal design requires using interjections and inter/intra-sentence break times in order to enhance the affective content of the synthesized speech. Besides, the gesture generator synthesizes adaptive head-arm gestures to the generated speech. The proposed design of facial expressions requires using additional body gestures in order to increase their credibility and expressivity to the participants.

This paper validates the important role of the robot behavioral multimodality in enhancing the clarity of interaction compared to interaction conditions with less affective cues. Moreover, it discusses the positive effect of the designed facial expressions and gestures in enhancing the emotional expressivity and recognizability of the robot behavior. Over and above, it demonstrates the perceptual differences between the introverted-extroverted and male-female participants for the generated affective robot behavior. For the future work, we are considering to improve the gestural expressivity of the system through additional gesture generators. Moreover, we are considering to ameliorate the affective expressivity of speech and facial expressions in order to make the generated multimodal robot behavior more persuasive and natural. Besides, we are considering to integrate language models that can help the robot to under-

stand human language with a wider scope instead of parsing keywords as with the employed system in the paper [1, 8–11].

8 Conflict of Interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Aly A, Taniguchi T (2018) Towards understanding language through perception in situated human-robot interaction: From word grounding to grammar induction. In: Proceedings of the International Conference on Social Cognition in Humans and Robots (socSMCs), Hamburg, Germany 24
2. Aly A, Tapus A (2011) A voice-based gender and internal state combined detection model. In: Proceedings of the 6th ACM/IEEE Human-Robot Interaction Conference (HRI), Switzerland 3
3. Aly A, Tapus A (2012) A model for mapping speech to head gestures in human-robot interaction. In: Borangiu T, Thomas A, Trentesaux D (eds) Service Orientation in Holonic and Multi-Agent Manufacturing Control, Studies in Computational Intelligence, Springer, Heidelberg, pp 183–196 2
4. Aly A, Tapus A (2013) Prosody-based adaptive metaphoric head and arm gestures synthesis in human robot interaction. In: Proceedings of the 16th IEEE International Conference on Advanced Robotics (ICAR), Montevideo, Uruguay, pp 1–8 2, 4, 7, 9, 14
5. Aly A, Tapus A (2014) Towards enhancing human-robot relationship: Customized robot's behavior to human's profile. In: Proceedings of the AAAI Fall Symposium on AI for Human-Robot Interaction (AI-HRI), VA, USA 2
6. Aly A, Tapus A (2015) Multimodal adapted robot behavior synthesis within a narrative human-robot interaction. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hamburg, Germany, pp 2986–2993 15
7. Aly A, Tapus A (2016) Towards an intelligent system for generating an adapted verbal and nonverbal combined behavior in human-robot interaction. Autonomous Robots 40(2):193–209 4
8. Aly A, Taniguchi T, Mochihashi D (2018) A Bayesian approach to phrase understanding through cross-situational learning. In: International Workshop on Visually Grounded Interaction and Language (ViGIL), in Conjunction with the 32nd Conference on Neural Information Processing Systems (NeurIPS), Montreal, Canada 24
9. Aly A, Taniguchi T, Mochihashi D (2018) A probabilistic approach to unsupervised induction of combinatory categorial grammar in situated human-robot interaction. In: Proceedings of the 18th IEEE-RAS International Conference on Humanoid Robots (Humanoids), Beijing, China, pp 1–9
10. Aly A, Taniguchi T, Mochihashi D (2018) Towards understanding syntactic structure of language in human-robot interaction. In: International Workshop on Visually Grounded Interaction and Language (ViGIL), in Conjunction with the

32nd Conference on Neural Information Processing Systems (NeurIPS), Montreal, Canada
11. Aly A, Jenkins OC, Sabanovic S (2019) Representation learning in HRI. ACM Transactions on Human-Robot Interaction (THRI) 8(4) 24
12. Atta M, Ather M, Bano M (2013) Emotional intelligence and personality traits among university teachers: Relationship and gender differences. International Journal of Business and Social Science 4(17):253–259 20
13. Beira R, Lopes M, Praga M, Santos-Victor J, Bernardino A, Metta G, Becchi F, Saltaren R (2006) Design of the robot-cub (iCub) head. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), USA, pp 94–100 2, 5
14. Boden MT, Thompson RJ, Dizén M, Berenbaum H, Baker JP (2013) Are emotional clarity and emotion differentiation related? Cognition and Emotion 27(6):961–978 12
15. Breazeal C (2003) Towards sociable robots. Robotics and Autonomous Systems 42:167–175 5
16. Breemen AV, Yan X, Meerbeek B (2005) iCat: An animated user-interface robot with personality. In: Proceedings of the 4th International Conference on Autonomous Agents and Multiagent Systems (AAMAS), Utrecht, Netherlands 5
17. Busso C, Deng Z, Yildirim S, Bulut M, Lee C, Kazemzadeh A, Lee S, Neumann U, Narayanan S (2004) Analysis of emotion recognition using facial expressions, speech, and multimodal information. In: Proceedings of the 6th International Conference on Multimodal Interfaces (ICMI), NY, USA, pp 205–211 2, 5
18. Caridakis G, Castellano G, Kessous L, Raouzaiou A, Malatesta L, Asteriadis S, Karpouzis K (2007) Multimodal emotion recognition from expressive faces, body gestures and speech. In: Boukis C, Pnevmatikakis A, Polymenakos L (eds) Artificial Intelligence and Innovations 2007: From Theory to Applications AIAI, IFIP The International Federation for Information Processing, vol 247, Springer, Boston, MA 2
19. Cassell J, Bickmore T, Campbell L, Vilhjálmsson H, Yan H (2000) Human conversation as a system framework: Designing embodied conversational agents. In: Cassell J, Sullivan J, Prevost S, Churchill E (eds) Embodied Conversational Agents, MIT Press, MA, USA, pp 29–63 4
20. Cassell J, Vilhjálmsson HH, Bickmore T (2001) BEAT: The behavior expression animation toolkit. In: Proceedings of the SIGGRAPH, pp 477–486 2, 4
21. Castellano G, Kessous L, Caridakis G (2007) Emotion recognition through multiple modalities: Face, body gesture, speech. In: Peter C, Beale R (eds) Affect and Emotion in Human Computer Interaction, Lecture Notes in Computer Science, vol 4868, Springer, Heidelberg 6, 22
22. Chiu CC, Morency LP, Marsella S (2015) Predicting co-verbal gestures: A deep and temporal modeling approach. In: Proceedings of the ACM International Conference on Intelligent Virtual Agents (IVA), pp 152–166 4
23. Clavel C, Plessier J, Martin JC, 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 (IVA), pp 287–300 5

24. Coffey E, Berenbaum H, Kerns JG (2003) The dimensions of emotional intelligence, alexithymia, and mood awareness: Associations with personality and performance on an emotional stroop task. Cognition and Emotion 17(4):671–679 12, 13
25. Costa S, Soares F, Santos C (2013) Facial expressions and gestures to convey emotions with a humanoid robot. In: Herrmann G, Pearson MJ, Lenz A, Bremner P, Spiers A, Leonards U (eds) Social Robotics (ICSR), Lecture Notes in Computer Science, vol 8239, Springer, pp 542–551 5
26. Cowie R, Cornelius R (2003) Describing the emotional states that are expressed in speech. Speech Communication 40:5–32 3
27. Edgington M (1997) Investigating the limitations of concatenative synthesis. In: Proceedings of Eurospeech, Greece 3
28. Ekman P (1979) About brows: Emotional and conversational signal. In: von Cranach M, Foppa K, Lepenies W, Ploog D (eds) Human Ethology: Claims and Limits of a New Discipline: Contributions to the Colloquium, Cambridge University Press, Cambridge, UK, pp 169–248 2
29. Ekman P, Friesen WV (1969) The repertoire of nonverbal behavior: Categories, origins, usage, and coding. Semiotica 1:49–98 3
30. Ekman P, Friesen WV (1978) Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press, CA, USA 8
31. Gibiansky A, Arik S, Diamos G, Miller J, Peng K, Ping W, Raiman J, Zhou Y (2017) Deep voice 2: Multi-speaker neural text-to-speech. In: Proceedings of the International Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, USA, pp 2962–2970 3
32. Goldberg LR (1990) An alternative description of personality: The Big-Five factor structure. Personality and Social Psychology 59:1216–1229 20
33. Gunes H, Celiktutan O, Sariyanidi E (2019) Live human–robot interactive public demonstrations with automatic emotion and personality prediction. Philosophical Transactions of the Royal Society B 374(1771) 2
34. Hasegawa D, Kaneko N, Shirakawa S, Sakuta H, Sumi K (2018) Evaluation of speech-to-gesture generation using bi-directional LSTM network. In: Proceedings of the ACM International Conference on Intelligent Virtual Agents (IVA), Sydney, NSW, Australia 4
35. Hermans D, Houwer JD, Eelen P (2001) A time course analysis of the affective priming effect. Cognition and Emotion 15(2):143–165 17
36. Hewig J, Hagemann D, Seifert J, Gollwitzer M, Naumann E, Bartussek D (2005) A revised film set for the induction of basic emotions. Cognition and Emotion 19(7):1095–1109 2, 6, 12, 16, 17
37. Hoffman G, Zuckerman O, Hirschberger G, Luria M, Shani-Sherman T (2015) Design and evaluation of a peripheral robotic conversation companion. In: Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Portland, USA 5
38. Hortensius R, Hekele F, Cross ES (2018) The perception of emotion in artificial agents. IEEE Transactions on Cognitive and Developmental Systems 10(4):852–864 5, 6, 22

39. Hutcherson CA, Goldin PR, Ramel W, McRae K, Gross JJ (2008) Attention and emotion influence the relationship between extraversion and neural response. Social Cognitive and Affective Neuroscience 3(1):71–79 20
40. Iida A, Campbell N (2003) Speech database design for a concatenative text-to-speech synthesis system for individuals with communication disorders. Speech Technology 6(4):379–392 2, 3
41. Kalra P, Mangili A, Magnenat-Thalmann N, Thalmann D (1991) SMILE: A multilayered facial animation system. In: Kunii T (ed) Modeling in Computer Graphics, Springer-Verlag, pp 189–198 4
42. Kang SM, Shaver PR (2004) Individual differences in emotional complexity: Their psychological implications. Personality 72(4):687–726 12
43. Karras T, Aila T, Laine S, Herva A, Lehtinen J (2017) Audio-driven facial animation by joint end-to-end learning of pose and emotion. ACM Transactions on Graphics 36(4) 5
44. Kashdan TB, Barrett LF, McKnight PE (2015) Unpacking emotion differentiation: Transforming unpleasant experience by perceiving distinctions in negativity. Current Directions in Psychological Science 24(1):10–16 13
45. Kendon A (1983) The study of gesture: Some remarks on its history. In: Deely J, Lenhart M (eds) Semiotics 1981, Springer-Verlag, NY, USA, pp 153–164 3
46. Kendon A (1988) How gestures can become like words. In: Poyatos F (ed) Cross Cultural Perspectives in Non-Verbal Communication, Hogrefe, Toronto, Canada, pp 131–141 3
47. Kim H, Lu X, Costa M, Kandemir B, Adams RB, Li J, Wang JZ, Newman MG (2018) Development and validation of image stimuli for emotion elicitation (ISEE): A novel affective pictorial system with test-retest repeatability. Psychiatry Research 261(414-420) 15
48. Kopp S, Wachsmuth I (2004) Synthesizing multimodal utterances for conversational agents. Computer Animation and Virtual Worlds 15(1):39–52 4
49. Kucherenko T, Hasegawa D, Henter GE, Kaneko N, Kjellstrom H (2019) Analyzing input and output representations for speech-driven gesture generation. In: Proceedings of the ACM International Conference on Intelligent Virtual Agents (IVA), Paris, France 4
50. Le QA, Huang J, Pelachaud C (2012) A common gesture and speech production framework for virtual and physical agents. In: Proceedings of the 14th ACM International Conference on Multimodal Interaction (ICMI), CA, USA 4
51. Lee Y, Rabiee A, Lee SY (2017) Emotional end-to-end neural speech synthesizer. In: Proceedings of the International Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, USA 3
52. Lutkebohle I, Hegel F, Schulz S, Hackel M, Wrede B, Wachsmuth S, Sagerer G (2010) The Bielefeld anthropomorphic robot head Flobi. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), AK, USA, pp 3384–3391 5
53. Mayer JD, Salovey P (1997) What is emotional intelligence? In: Salovey P, Sluyter D (eds) Emotional Development and Emotional Intelligence: Educational Implications, Basic Books, NY, USA, pp 3–34 20

54. Mayer JD, Roberts RD, Barsade SG (2008) Human abilities: Emotional intelligence. Annual Review of Psychology 59:507–536 13
55. McColl D, Nejat G (2014) Recognizing emotional body language displayed by a human-like social robot. International Journal of Social Robotics 6:261–280 9
56. McNeill D (1992) Hand and Mind: What Gestures Reveal about Thought. University of Chicago Press, IL, USA 3, 4, 14
57. Mozziconacci S (2002) Prosody and emotions. In: Proceedings of the International Conference on Speech Prosody, Aix-en-Provence, France, pp 1–9 4
58. Murray IR, Arnott JL (1995) Implementation and testing of a system for producing emotion-by-rule in synthetic speech. Speech Communication 16(4):369–390 2, 3
59. Oudeyer PY (2003) The production and recognition of emotions in speech: Features and algorithms. Human-Computer Studies 59(1):157–183 3
60. Park E, Kim KJ, del Pobil AP (2011) The effects of robot's body gesture and gender in human-robot interaction. In: Proceedings of the 15th International Conference on Internet and Multimedia Systems and Applications, Washington DC., USA 21
61. Pelachaud C (2005) Multimodal expressive embodied conversational agents. In: Proceedings of the 13th Annual ACM International Conference on Multimedia, NY, USA, pp 683–689 2, 4
62. Pell MD, Kotz SA (2011) On the time course of vocal emotion recognition. PLoS ONE 6(11) 3
63. Petrides KV, Vernon PA, Schermer JA, Ligthart L, Boomsma DI, Veselka L (2010) Relationships between trait emotional intelligence and the Big Five in the Netherlands. Personality and Individual Differences 48:906–910 20
64. Picard RW (2003) Affective computing: Challenges. International Journal of Human-Computer Studies 59(1–2):55–64 5
65. Platt SM, Badler N (1981) Animating facial expressions. Computer Graphics 15:245–252 2, 5
66. Qian Y, Fan Y, Hu W, Soong FK (2014) On the training aspects of deep neural network DNN for parametric TTS synthesis. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 3829–3833 3
67. Reisenzein R, Weber H (2009) Personality and emotion. In: Corr P, Matthews G (eds) The Cambridge Handbook of Personality-Psychology, Cambridge University Press, Cambridge, UK, pp 54–71 20
68. Revelle W, Scherer KR (2010) Personality and emotion. In: Sander D, Scherer K (eds) The Oxford Companion to Emotion and the Affective Sciences, Oxford University Press, Oxford, UK 20
69. Ribeiro FS, Santos FH, Albuquerque PB, Oliveira-Silva P (2019) Emotional induction through music: Measuring cardiac and electrodermal responses of emotional states and their persistence. Frontiers in Psychology 15
70. Roberts NA, Tsai JL, Coan JA (2007) Emotion elicitation using dyadic interaction tasks. In: Coan JA, Allen JJB (eds) Handbook of Emotion Elicitation and Assessment, Series in Affective Science, Oxford University Press 15

71. Salem M, Rohlfing K, Kopp S, Joublin F (2011) A friendly gesture: investigating the effect of multimodal robot behavior in human–robot interaction. In: Proceedings of the 20th IEEE International Symposium on Robot and Human Interaction Communciation (RO-MAN), pp 247–252 5
72. Sauter DA, Eisner F, Calder AJ, Scott SK (2010) Perceptual cues in nonverbal vocal expressions of emotion. The Quarterly Journal of Experimental Psychology 63(11):2251–2272 3
73. Schirmer A, Adolphs R (2017) Emotion perception from face, voice, and touch: Comparisons and convergence. Trends in Cognitive Sciences 21(3):216–228 6, 23
74. Schreuder E, Erp JV, Toet A, Kallen VL (2016) Emotional responses to multisensory environmental stimuli: A conceptual framework and literature review. SAGE Open 6:1–19 2, 14, 17
75. Schroder M, Trouvain J (2003) The German text-to-speech synthesis system Mary: A tool for research, development, and teaching. Speech Technology 6(4):365–377 3, 7
76. Shichuan D, Yong T, Martinez AM (2014) Compound facial expressions of emotion. In: Proceedings of the National Academy of Sciences of the United States of America (PNAS), vol 111, pp 1454–1462 9
77. Shulman TE, Hemenover SH (2006) Is dispositional emotional intelligence synonymous with personality? Self and Identity 5(2):147–171 20
78. Siegel M, Breazeal C, Norton M (2009) Persuasive robotics: The influence of robot gender on human behavior. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), MO, USA, pp 2563–2568 21
79. Silva LCD, Miyasato T, Nakatsu R (1997) Facial emotion recognition using multimodal information. In: Proceedings of IEEE International Conference on Information, Communications, and Signal Processing (ICICS), Singapore, vol 1, pp 397–401 5
80. Spencer-Smith J, Wild H, Innes-Ker A, Townsend JT, Duffy C, Edwards C, Ervin K, Merritt N, Paik JW (2001) Making faces: Creating three-dimensional parameterized models of facial expression. Behavior Research Methods, Instruments, and Computers 33(2):115–123 5
81. Tapus A, Aly A (2011) User adaptable robot behavior. In: Proceedings of the IEEE International Conference on Collaboration Technologies and Systems (CTS), PA, USA 2
82. Taylor P, Isard A (1997) SSML: A speech synthesis markup language. Speech Communication 21:123–133 7
83. Taylor S, Kim T, Yue Y, Mahler M, Krahe J, Rodriguez AG, Hodgins J, Matthews I (2017) A deep learning approach for generalized speech animation. ACM Transactions on Graphics 36(4) 5
84. Tsiourti C, Weiss A, Wac K, MVincze (2019) Multimodal integration of emotional signals from voice, body, and context: Effects of (in)congruence on emotion recognition and attitudes towards robots. International Journal of Social Robotics 11:555–573 5

85. Uhrig MK, Trautmann N, Baumgartner U, Treede RD, Henrich F, Hiller W, Marschall S (2016) Emotion elicitation: A comparison of pictures and films. Frontiers in Psychology 15
86. Um SY, Oh S, Byun K, Jang I, Ahn C, Kang HG (2020) Emotional speech synthesis with rich and granularized control. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain 3
87. Vlachos E, Schärfe H (2012) Android emotions revealed. In: Proceedings of the 4th International Conference on Social Robotics (ICSR), Chengdu, China, pp 56–65 5
88. Vougioukas K, Petridis S, Pantic M (2018) End-to-end speech-driven facial animation with temporal GANs. In: Proceedings of the British Machine Vision Conference (BMVC), UK 5
89. Wallbott HG (1998) Bodily expression of emotion. European Journal of Social Psychology 28:879–896 9
90. Wang Y, Skerry-Ryan RJ, Stanton D, Wu Y, Weiss RJ, Jaitly N, Yang Z, Xiao Y, Chen Z, Bengio S, Le Q, Agiomyrgiannakis Y, Clark R, Saurous RA (2017) Tacotron: Towards end-to-end speech synthesis. In: Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH) 3
91. Yoon Y, Ko WR, Jang M, Lee J, Kim J, Lee G (2019) Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In: Proceedings of the International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, pp 4303–4309 4
92. Zen H, Senior A, Schuster M (2013) Statistical parametric speech synthesis using deep neural networks. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 7962–7966 3
93. Zeng Z, Pantic M, Roisman GI, Huang TS (2009) A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(1) 22