2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
Congress Center Hamburg
Sept 28 - Oct 2, 2015. Hamburg, Germany

Multimodal Adapted Robot Behavior Synthesis within a Narrative Human-Robot Interaction

Amir Aly[1] and Adriana Tapus[2]

Abstract— **In human-human interaction, three modalities of communication (i.e., verbal, nonverbal, and paraverbal) are naturally coordinated so as to enhance the meaning of the conveyed message. In this paper, we try to create a similar coordination between these modalities of communication in order to make the robot behave as naturally as possible. The proposed system uses a group of videos in order to elicit specific target emotions in a human user, upon which interactive narratives will start (i.e., interactive discussions between the participant and the robot around each video's content). During each interaction experiment, the humanoid expressive ALICE robot engages and generates an adapted multimodal behavior to the emotional content of the projected video using speech, head-arm metaphoric gestures, and/or facial expressions. The interactive speech of the robot is synthesized using Mary-TTS (text to speech toolkit), which is used - in parallel - to generate adapted head-arm gestures [1]. This synthesized multimodal robot behavior is evaluated by the interacting human at the end of each emotion-eliciting experiment. The obtained results validate the positive effect of the generated robot behavior multimodality on interaction.**

I. INTRODUCTION

The need for an intelligent robot that can customize the emotional content of its synthesized multimodal behavior to the context of interaction so as to increase the credibility of its communicative intents, is increasing rapidly. Speech, gestures, and facial expressions are used together to convey coordinated and synchronized verbal, paraverbal, and nonverbal information that could enhance the content of interaction. The importance of gestures and facial expressions lies in their ability to clarify the meaning of speech when the signal is deteriorated, in addition to the fact that they can replace or accompany words in a synchronized manner [2].

The correlation between emotion and speech had been intensively investigated in the literature. Speech prosody can reflect human emotion through changes in basic cues, such as: pitch, intensity, rate, and pauses [3]. The variation in the characteristics of voice prosody for different emotions: anger, disgust, fear, and sadness, was studied in [4]. The process of emotion perception and decoding, in addition to the required time to recognize different emotions based on their prosodic cues, was studied in [5]. The evolutionary nature of emotion was considered in [6] and [7], while studying the cognitive perception of emotion through a fuzzy model.

[1]Amir Aly is a postdoctoral research fellow in the Robotics and Computer Vision lab. in ENSTA-ParisTech, 828 Boulevard des Maréchaux, 91120 Palaiseau, France amir.aly@ensta-paristech.fr

[2]Prof. Adriana Tapus is with the Robotics and Computer Vision lab. in ENSTA-ParisTech, 828 Boulevard des Maréchaux, 91120 Palaiseau, France adriana.tapus@ensta-paristech.fr

On the way towards synthesizing emotional speech that can add more naturalness to human-robot interaction and human-computer interaction, the first emotional speech synthesis system was developed based on the rule-based formant synthesis technique [8], but the quality was a bit poor. Another interesting approach based on the diphone concatenation technique that achieved some limited success in expressing specific emotions, was discussed in [9]. This last technique was later developed to the unit selection technique that tries to avoid interference with the recorded voice during synthesis so as to obtain a better quality, and reported some small success in expressing only three emotions: happiness, anger, and sadness [10]. Generally, the previously discussed techniques are missing some explicit control on the prosodic parameters of speech so as to be able to express emotions on a wider scope. This constraint and the quality of the synthesized voice, constituted our inspiration for using a more controllable and efficient text-to-speech engine, like Mary-TTS [11], in our work.

On the other hand, the basic definition for gesture was given by Kendon [12] and McNeill [13]. They defined a gesture as a synchronized body movement with speech, which is related parallely or complementarily to the meaning of the utterance. The first step towards categorizing gestures, was discussed in Ekman and Friesen [14]. They proposed 5 gesture categories: (1) emblems, (2) illustrators, (3) facial expressions, (4) regulators, and (5) adaptors. However, Kendon [15] criticized the proposed gesture categories of Ekman for ignoring the linguistic phenomena. Therefore, he proposed a new classification for gestures of 4 categories: (1) gesticulation, (2) pantomime, (3) emblem, and (4) signs. McNeill [13] presented a more elaborate widely used gesture classification of 4 categories: (1) iconics (i.e., gestures representing images of concrete entities), (2) metaphorics (i.e., gestures representing abstract ideas), (3) deictics (i.e., pointing), and (4) beats (i.e., rhythmic finger movements).

Several research studies in the fields of human-robot interaction and human-computer interaction, have focused on synthesizing iconic and metaphoric gestures, which form together (according to McNeill) the major part of the generated nonverbal behavior during human-human interaction. Pelachaud [16] developed the 3D agent GRETA, which can synthesize a multimodal synchronized behavior to the human users based on an input text. Generally, GRETA can synthesize gestures of all categories regardless of the domain of interaction, to the contrary of other 3D conversational agents (e.g., the conversational agent MAX) [17]. An interesting framework was discussed in [18], which can synthesize a

multimodal synchronized behavior for both the 3D agent GRETA and robots. Cassell et al., [19] presented BEAT toolkit, which is a rule-based gesture generator. It applies the natural language processing (NLP) algorithms on an input text in order to produce an animation script that can animate both of humanoid robots [20], [21], and virtual agents (e.g., REA agent) [22]. This toolkit can generate gestures of different kinds (including iconic gestures) except for metaphoric gestures. Generally, the majority of gesture generation approaches are not considering the effect of emotion (expressed through prosody) on body language, which puts a difficulty towards adapting the generated robot behavior to human emotion [23]. In this paper, we present an ***extension*** of our previous work [1], [24], which proposes a statistical model for synthesizing adapted head-arm metaphoric gestures to the prosodic cues of speech *(for this reason, the experimental design in this paper is based on a human-robot interaction scenario in order to use the modeled synthesized speech on the robot (whose content corresponds to the comment of the interacting human on the video) as an input to the gesture generator)*. This model has been integrated to the system of this paper in order to generate an adapted multimodal robot behavior to the emotional content of interaction.

On the other hand, the correlation between facial expressions and speech had been long recognized in psychological studies [25]. The movement of face muscles and the prosodic cues of speech can change in a synchronized manner in order to communicate different emotions. The single-modal based perception of human emotion through audio or visual information, was discussed in [26]. Chen et al., [27] discussed the complementarity of both modalities, so that the perception of human emotion will be ameliorated when both modalities are considered in the same time. These last findings are considered in the experimental design of our study.

The synthesis and modeling of facial expressions in computer-based applications and 3D agents received more attention than in human-robot interaction. Parke [28] developed the first 3D face model that can convey different expressions. Platt and Badler [29] presented the first model that employs FACS (Facial Action Coding system) in controlling the muscular actions corresponding to facial expressions. Spencer-Smith et al., [30] developed a more 3D realistic model that can create a stimuli with 16 different FACS action units and determined intensities. Similarly, robots underwent different studies aiming towards allowing them to generate reasonable facial expressions. An early initiative to model facial expressions on robots was taken by Breazeal [31], who developed the robot head Kismet. It uses facial details, like: eyes, mouth, and ears to model facial expressions, such as: anger, happiness, surprise, sadness, and disgust. Breemen et al., [32] developed the research platform iCat, which can render different facial expressions, such as: sadness, anger, happiness, and fear. Beira et al., [33] developed the complete expressive humanoid robot iCub, which can synthesize a variety of emotions using gestures and facial expressions, including: anger, sadness, surprise, and happiness. In this paper, we use the highly expressive ALICE robot (Section II-C) for the purpose of generating a complete multimodal robot behavior, which was not sufficiently addressed before in the literature. The rest of the paper is structured as following: Section (II) presents a detailed illustration for the system architecture, Section (III) illustrates the design, the hypotheses, and the scenario of interaction, Section (IV) provides a description for the experimental results, and finally, Section (V) concludes the paper.

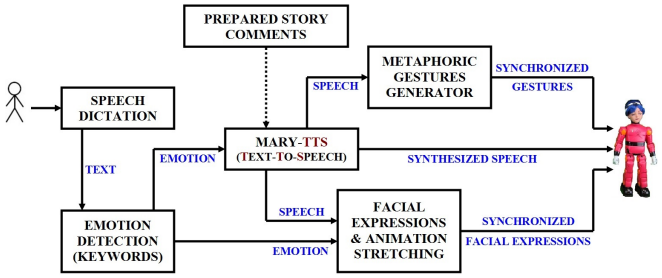


Fig. 1: Overview of the emotionally-adapted narrative system architecture

II. SYSTEM ARCHITECTURE

The proposed system is coordinated through different subsystems: (1) Speech dictation system (HTML5 API *multilingual* dictation toolkit), (2) Emotion detection phase, in which some defined keywords are parsed from the dictated speech of the human user so as to precise an emotional label for the video's content, (3) Mary-TTS engine, which converts the prepared texts (i.e., robot comments) and the detected emotion in each interaction to a synthesized emotional speech, (4) Metaphoric gestures generator, which maps the synthesized speech to synchronized head-arm metaphoric gestures [1], (5) Facial expressions modeling and animation stretching phase, and finally (6) ALICE robot as the test-bed platform in the experiments (Section III). An overview of the system architecture is illustrated in Figure (1).

A. Metaphoric Gestures Generator

The generator uses the Coupled (i.e., 2 chains for speech and gestures) Hidden Markov Models (CHMM) [34] in order to synthesize head-arm metaphoric gestures, as illustrated in details in [1] and [35]. The training of the CHMM requires first segmenting the characteristic curves of speech and gestures. The motion curves of gestures (i.e., the velocity, acceleration, displacement, and position curves) are segmented by calculating the force, momentum, and kinetic energy of body segments (e.g., the up-arm, low-arm, and hand segments), in addition to the total force of the body. The intersection between these descriptors represents the boundary points of gestures in each body segment. Meanwhile, the pitch-intensity curves of speech are segmented in parallel with gestures in terms of the boundary points of each gesture, the frame time, and the sampling frequency [1]. These segmented patterns of speech and gestures are used to train the CHMM, through which new adapted head-arm metaphoric gestures will be synthesized based on the prosodic cues of a speech-test signal.

```
<?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

B. Affective Speech Synthesis

The text-to-speech engine (Mary-TTS) is used for the purpose of adding relevant prosodic and accent cues to a pre-prepared text that summarizes the content of the video under discussion [11]. This allows the robot to engage in the conversation using - to some extent - adapted emotional speech to the emotional context of the video. The designed vocal patterns are represented using a low-level markup language called MaryXML (which is based on the XML markup language) or using other relatively high-level markup languages, like the SSML (Speech Synthesis Markup Language) [36]. The SSML representation offers more vocal design features, like imposing a silence period between words, in addition to an easy control on the characteristics of the pitch contour, baseline pitch, and speech rate (Figure 2), which makes it helpful for the emotional vocal patterns' design described in this study. However, the fact that Mary-TTS engine is not prepared yet for efficiently synthesizing emotional speech of different classes in the English language *(to the best of our knowledge, no other vocal engine can)*, makes our proposed vocal design as an *approximate* step towards conveying - even to some extent - the true meaning of the expressed emotion to a human user. Therefore, the multimodality of the robot behavior is a good solution that could emphasize the meaning of the expressed behavior, so that each modality enhances the other one.

The designed vocal patterns of the target emotions are summarized in Table (I), in which the pitch contours are characterized by sets of parameters inside parentheses (where the first parameter in each set followed by *"%"* represents a percentage of the text duration, while the second parameter represents the corresponding change in the baseline pitch in semitone, which is half of a tone on the standard diatonic scale). The speech rates of the target emotions vary between the rates of the “sadness" emotion (lowest rate) and the “anger" emotion (highest rate). The inter-sentence break time of each target emotion represents the silence periods between sentences (i.e., story texts), during which the robot's lips/jaw will make certain expressions that could enhance the expressed emotion (Section II-C). Meanwhile, the intra-sentence break time represents the short silence periods within a sentence, which are necessary for increasing the credibility of the “sadness" and “fear" emotions.

The indicated experimental parameters in Table (I) give an example to the prosodic patterns of parts of the texts that Mary-TTS engine should convert to speech in different emotions. The other prosodic patterns of the remaining parts of the texts could differ slightly from the contour parameters of Table (I) in order to show some tonal variation through the total of each text.

C. Face Expressivity

The designed facial expressions corresponding to the prescribed target emotions in this study, are based on the Facial Action Coding System (FACS) [37]. Table (II) illustrates the FACS coding of each target emotion in this study, in addition to the available equivalent joints in the face of the robot that we used in order to model each expression in the most persuasive manner.

The complexity behind modeling emotions on the face of the robot lies in the absence of the equivalent joints to some FACS descriptors (e.g., cheek raiser, nose wrinkler). Therefore, we imposed experimentally some additional body gestures in order to reduce the negative effect of the missing joints so as to enhance the expressed emotion. These additional gestures do not include -normally- any head gesture (i.e., neck rotation) nor arm-hand gestures, which are being generated by the metaphoric gestures generator explained earlier (Section II-A).[1] However, the combination of the *neck rotation* (i.e., turning the head aside) and the *raising front-bent arms* has been helpful for better expressing the “disgust" emotion (consequently, they got considered as additional supportive gestures for this emotion). This will help give the interacting human - even to some extent - the impression that the robot did not like the context of interaction and considered it disgusting. Similarly, the “fear" emotion, the “anger" emotion, and the “sadness" emotion have been attributed additional *mouth-guard hand* gesture, *down head-shaking*, and *bowing head and covering-eyes hand* gesture respectively, in order to help emphasize their meanings, as indicated in Figure (3). On the other hand, the main role of the additional supportive *left smile* and *right smile* face joints of the “fear" emotion, is to depress a little the corners of the open mouth to better reflect the emotion, however they do not have any equivalent FACS descriptor representing the “fear" emotion, as indicated in Table (II).

Generally, the modeling of facial expressions on a humanoid robot (even with the expressive ALICE robot) is not an easy task due to the mechanical limitations that the robot has (unlike the 3D agents). Therefore, the multimodality of the robot behavior is important for interaction, which makes each modality of behavior expression enhances the other modalities so as to emphasize the conveyed meaning of the expressed emotion to a human user.

The temporal alignment between the synthesized emotional speech and the designed facial expressions is controlled by the duration of the generated speech. In case the

[1]The metaphoric gestures generator [1] has the liberty to synthesize the most appropriate gestures based on its own learning algorithm. Therefore, it is *probable* that the previously mentioned supportive head-arm gestures will not be synthesized by the generator during interaction. Consequently, we imposed them at specific moments during speech with a higher priority than the generator's synthesized gestures to make sure of their presence.

TABLE I: Approximate design of the vocal pattern and the corresponding contour behavior of each target emotion on the standard diatonic scale. Some emotions have used interjections (with tonal stress) in order to emphasize the desired meaning, like: 'Shit' for the "anger" emotion, 'Ugh' and 'Yuck' for the "disgust" emotion, and 'Oh my God' for the "fear" 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

TABLE II: FACS coding of the target emotions and the corresponding joints in the robot's face, in addition to the other required gestures to emphasize the meaning of facial expression. The bold FACS action units in each emotion represent the observed prototypical units between the subjects in [38], while the other less common non-bold units are observed with different lower percentages between the subjects. The underlined action units represent the units that have *approximate* corresponding joints in the robot's face.

Emotion	FACS Coding	Robot's Face Joint	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*

duration of the generated speech is longer or shorter than the preliminary duration of the animation, the system calculates easily the new time instant of each control point composing the XML animation script (in which the control points are characterized in terms of position and time) as a function of the new duration of the generated speech, the preliminary duration of the animation, and the last time instant value of each control point. Meanwhile, the position of each control point is kept unchanged during the animation.

The segmentation of human speech employs the voice activity detection algorithm in order to label and separate between the speech and silence segments. In case the silence period is related to an inter-sentence break time (Section II-B), the robot's lips/jaw perform certain expressions (e.g., lip corner pulling for the "happiness" emotion) to enhance the conveyed meaning of the expressed emotion, as indicated in Figure (3). This is due to the mechanical limitations of the robot that do not allow for synchronizing the lips with speech, while performing an expression with the lips/jaw in the same time. However, in case the silence period is related to an intra-sentence break time (Section II-B), the robot's jaw is kept opened in the "fear" emotion and closed in the "sadness" emotion, during the duration of the silence period.

On the other hand, the animation of the robot's lips in a synchronized manner with the segmented speech has encountered a big difficulty when using the 3 servo motors controlling the lips motion (2 motors for the corners and 1 motor for the vertical motion), because they can not generate a reasonable homogeneous motion when operating together during continuous speech, in addition to the noise they generate. Alternatively, we used only the motor that controls the vertical motion of the lips. Afterwards, the remote running server of the robot maps the calculated visemes corresponding to the segmented speech to lips motion.

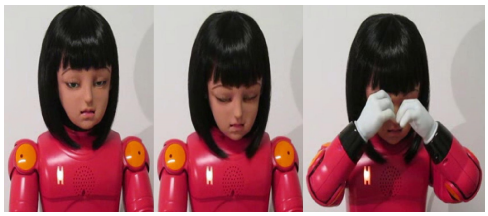
III. EXPERIMENTAL SETUP

In this section, we introduce the database used in inducing emotions in each participant, the experimental hypotheses, the design, and the scenario of interaction between the participant and ALICE robot developed by Hanson Robotics.[2]

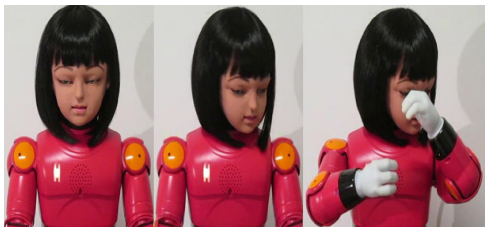
A. Database

The database used in this research contains 20 videos inducing the following 6 emotions: sadness, disgust, happiness, anger, fear, and neutral. The duration of the videos varies from 29 to 236 seconds, and all of them have been extracted from commercial films. The procedures of validating the efficiency of the database in eliciting the target emotions in

[2] The humanoid ALICE-R50 robot has a full-motion body and an expressive face, with a total of 36 degrees of freedom. The robot is equipped with two cameras and an array of sensors, including an accelerometer sensor, a torque sensor, a series of touch sensors, in addition to many other different sensors that allow it to precisely perceive its surroundings. The face of the robot composed of synthetic skin, is its main specialty. It can create a full range of credible facial expressions in different emotions (Section II-C).

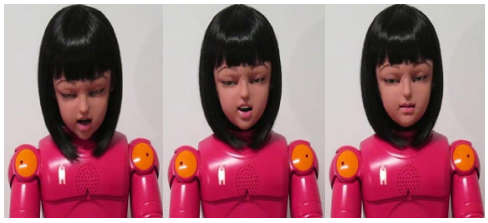


(a) Sadness

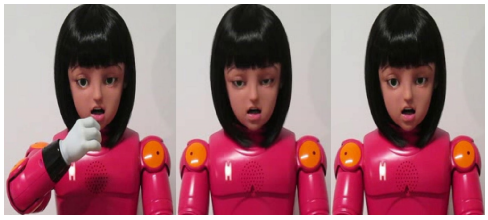


(b) Disgust

(c) Happiness



(d) Anger



(e) Fear

Fig. 3: Synthesized facial expressions by ALICE robot

humans, were discussed in [39]. During the experiments, we used 12 videos extracted from different films for eliciting the target emotions (which constitute 6 main videos used during the experiments, and 6 standby videos used automatically when the main videos fail to elicit the corresponding target emotions), as indicated in Table (III).

B. Hypotheses

This study aims to test and to validate the following hypotheses:

- **H**1: The combination of facial expressions, head-arm metaphoric gestures, and synthesized emotional speech will make the emotional content of interaction more clear to the participant than the interaction conditions that employ less affective cues.
- **H**2: Facial expressions will enhance the expressiveness of the robot emotion in contrast to the interaction conditions that do not employ facial expressions.
- **H**3: The dynamic characteristics of the robot head-arm metaphoric gestures will help the participant recognize and distinguish between the target emotions.

TABLE III: Target emotions and their corresponding feature films. The main videos used during the experiments were extracted from the bold feature films.

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

C. Experimental Design

Our design contains three robot conditions:

- The robot generates a combined multimodal behavior expressed through synchronized head-arm metaphoric gestures, facial expressions, and speech (i.e., condition C1-SFG).
- The robot generates a combined multimodal behavior expressed through synchronized facial expressions and speech (i.e., condition C2-SF).
- The robot generates a combined multimodal behavior expressed through synchronized head-arm metaphoric gestures and speech (i.e., condition C3-SG).
- The robot generates a single-modal behavior expressed only through speech (i.e., condition C4-S).

In order to examine the first hypothesis, the first three conditions were examined (in which the facial expressions are accompanied by the additional supportive gestures illustrated in Table II). In this hypothesis, we excluded the conditions of the robot expressing a single-modal behavior only through head-arm metaphoric gestures or facial expressions without speech, in addition to the condition of the robot expressing combined facial expressions and head-arm metaphoric gestures without speech, because they do not match the context of the ***non-mute*** human-human interaction. Consequently, the importance of speech in recognizing emotions is measured directly through the questionnaire. On the other hand, in order to validate the second hypothesis, two conditions were investigated, which are the same as the conditions C2-SF and C4-S. Similarly, in order to validate the third hypothesis, two other conditions were tested, which are the same as the conditions C3-SG and C4-S (the condition C2-SF was excluded from validating the third hypothesis, because facial expressions were accompanied by the additional supportive gestures explained earlier).

Both of the robot and the interacting human follow a series of short videos through 6 experiments that mean to elicit 6

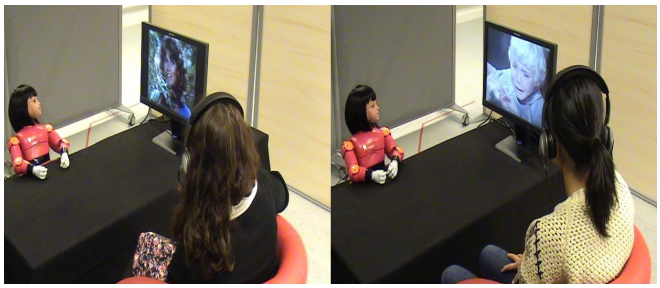


Fig. 4: Two participants interacting with the robot during the “happiness" and “sadness" emotion elicitation experiments

emotions (Section III-A) (Figure 4). The different methodologies of emotion induction and assessment were illustrated in [40]. The idea behind using videos to successfully elicit emotions in the human user, is that they are emotionally convincing and their role is well indicated in the literature [41]. An interesting study about eliciting emotions from films was discussed in [39], in which the results proved that the studied target emotions were reasonably recognized. In our study, the scenario of interaction is described as following:

- The robot welcomes the participant and invites him/her to watch some videos so as to have a discussion about.
- The robot asks the participant to express his/her opinion about the content of the video. Afterwards, it parses some expected emotional labels from the dictated comment of the participant, such as: This is *disgusting*!. This helps detect the video’s emotional content so as to trigger an adapted robot behavior.
- After listening to the comment of the participant on the video, the robot makes itself a comment accompanied by adapted emotional speech, head-arm metaphoric gestures, and/or facial expressions to the video’s content.
- In case the video rarely elicits a different emotion in the participant from the target emotion, so that the system parses some keywords that belong mainly to another category of non-target-emotion-referring keywords, the robot will comment through a *neutral* behavior in order to avoid any emotion-biasing effect. Then, it will invite the participant to watch another video, which means to certainly elicit the same emotion that was not successfully induced in the participant with the first video.
- The interaction ends for the concerned emotion. Afterwards, the participant starts evaluating the modeled behavior on the robot considering the relevance of its emotional content to the context of interaction, through a Likert questionnaire (in which all questions are presented on a 7-point scale). Whereupon, a new interaction for a new *randomly* selected emotion starts.
- After the experiments end up, the robot and the experimenter thank the participant for his/her cooperation.

IV. EXPERIMENTAL RESULTS

The experimental design was based on the between-subjects design, and 60 participant were recruited in order to validate our hypotheses. The participants were uniformly distributed between the four experimental conditions (15 participant (6 female and 9 male) /condition). The recruited participants were ENSTA-ParisTech undergraduate and graduate students and employees whose ages were varying between 20-57 years old ($M = 29.64$, $SD = 9.4$). The background of the participants was non-technical with an average of 33.3%, and technical with an average of 66.7%. 40% of the participants have interacted before with robots, while 60% of the participants have never interacted with robots beforehand.

For the first hypothesis, a significant difference was found by ANOVA analysis in the clearness of the adapted robot emotional behavior expressed through a combination of head-arm metaphoric gestures, facial expressions, and speech with respect to the robot emotional behavior expressed through facial expressions and speech, and the robot emotional behavior expressed through head-arm metaphoric gestures and speech ($F[2,267] = 9.69$, $p < 0.001$). Tukey’s HSD comparisons indicated a significant difference between the robot embodied with combined head-arm metaphoric gestures, facial expressions, and speech (i.e., condition C1-SFG) from one side, and the robot embodied with facial expressions and speech (i.e., condition C2-SF) ($p < 0.001$), in addition to the robot embodied with head-arm metaphoric gestures and speech (i.e., condition C3-SG) ($p < 0.001$) from the other side. However, no significant difference was observed between the experimental conditions C2-SF and C3-SG. Moreover, the participants found that the robot behavior was more expressive in the condition C1-SFG than in the condition C3-SG ($F[1,178] = 13.64$, $p < 0.001$). No significant differences were observed in the participants’ ratings regarding the naturalness of the robot behavior in the conditions C1-SFG, C2-SF, and C3-SG.

For the second hypothesis, the participants found that the robot behavior expressed though facial expressions and speech, was showing more expressiveness and was more adapted to the content of interaction than the robot behavior expressed only through speech (i.e., condition C4-S) ($F[1,178] = 16.27$, $p < 0.001$). Moreover, the participants considered that facial expressions and speech were synchronized with an average score of $M = 5.9$, $SD = 0.9$. Furthermore, they did not find any significant contradiction between the modalities of the robot behavior expressed through facial expressions and speech with an average score of $M = 1.8$, $SD = 1.2$. Over and above, they agreed that facial expressions were more expressive than speech with an average score of $M = 4.4$, $SD = 1.5$. Table (IV) shows that the facial expressions of the robot have only ameliorated the recognition score of the “anger" emotion in the condition C2-SF with respect to the condition C4-S. This amelioration is related to the encountered difficulties to design a highly persuasive vocal pattern for the “anger" emotion due to the limitations of the Mary-TTS engine (Section II-B). Therefore, the facial expressions of the robot have certainly enhanced the affective meaning of speech so as to give the participant a clear feeling that the robot was expressing the “anger" emotion. To the contrary, the facial expressions of the robot had a negative influence on the recognition score of

the “disgust" emotion in the condition C2-SF with respect to the condition C4-S, which is due to the limited expressivity of the robot’s face for this emotion (Section II-C).

TABLE IV: Recognition scores of the target emotions expressed by the robot in 3 different experimental 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%

For the third hypothesis, the participants considered that the emotional content of the robot behavior expressed through head-arm metaphoric gestures and speech was more observable than the emotional content of the robot behavior expressed only through speech ($F[1,178] = 17.16$, $p = 0.0001$). Furthermore, the participants found that gestures and speech were synchronized with an average score of $M = 6.1$, $SD = 0.7$. At the same time, they agreed that the execution of gestures was fluid with an average score of $M = 5.3$, $SD = 1.01$. Moreover, they considered that gestures were slightly more expressive than speech with an average score of $M = 4.2$, $SD = 1.4$. The emotional content of the robot head-arm metaphoric gestures was generally recognizable with reasonable scores, as indicated in Table (IV). However, they have only ameliorated the recognition score of the “anger" emotion in the condition C3-SG with respect to the condition C4-S (similarly to the effect of facial expressions), while the other recognition scores were equal in both conditions. Consequently, the dynamic characteristics of the generated gestures in case of the “anger" emotion, like the high velocity and acceleration, have certainly enhanced the expressive meaning of speech and gave the human the feeling that the robot was angry in a more persuasive manner.

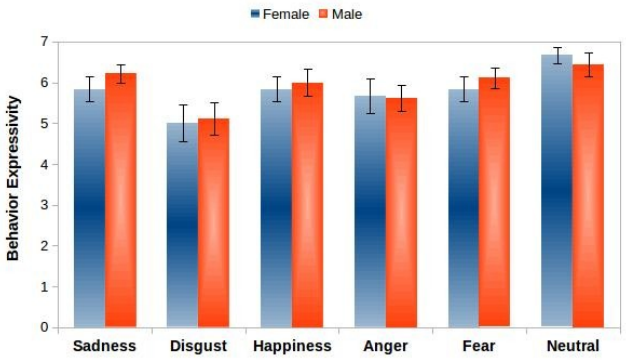


Fig. 5: Gender-based evaluation for the emotional expressiveness of the multimodal robot behavior expressed through combined facial expressions and speech (condition C2-SF). The error bars represent the calculated standard errors.

On the other hand, the emotional expressiveness of the robot behavior was positively perceived in general by the male and female participants in the conditions C2-SF and C3-SG, as indicated in Figures (5) and (6), respectively.

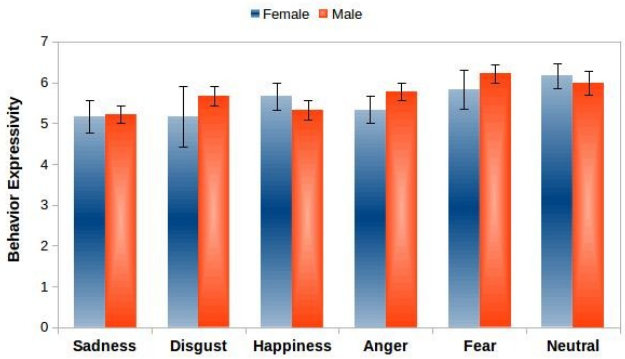


Fig. 6: Gender-based evaluation for the emotional expressiveness of the multimodal robot behavior expressed through combined head-arm gestures and speech (condition C3-SG). The error bars represent the calculated standard errors.

However, the perception of the male participants for the emotional expressiveness of the robot in both conditions, was generally higher than the perception of the female participants. The male participants in the condition C2-SF gave higher ratings for the emotions: sadness, disgust, happiness, and fear, meanwhile the female participants gave higher ratings for the emotions: anger and neutral (Figure 5). Similarly, the male participants in the condition C3-SG gave higher ratings for the emotions: sadness, disgust, anger, and fear, meanwhile the female participants gave higher ratings for the emotions: happiness and neutral (Figure 6). These findings reveal the relatively higher preference of the male participants for the emotional expressiveness of the female ALICE robot, than the female participants. This gender-based evaluation matches the findings of [42], which proved the tendency of the participants to consider the opposite-sex robots as being more credible, engaging, and persuasive.

V. CONCLUSIONS

This paper discusses adapting the multimodal robot behavior to the emotional content of a series of videos eliciting specific emotions in the human user within a narrative human-robot interaction. Each interacting human was exposed to only one of 4 different experimental conditions of multimodal/single-modal behaviors (i.e., conditions: C1-SFG, C2-SF, C3-SG, or C4-S) during the whole 6 experiments of eliciting the 6 target emotions. Our proposed system uses Mary-TTS engine in order to generate emotional speech. The metaphoric gestures generator synthesizes head-arm gestures based on the prosodic cues of speech. On the other hand, the designed facial expressions on the robot required some additional supportive gestures to enhance the conveyed meaning of the expressed emotion to the human.

This paper validates the role of the robot behavior multimodality in increasing the clearness of the emotional content of interaction with respect to the interaction conditions that use less affective cues. Besides, it proves the role of facial expressions in enhancing the expressiveness of the robot behavior, and the role of the generated gestures in

recognizing the target emotions. For the future work, we are interested in increasing the gestural expressivity of the system by integrating additional gesture generators, which can synthesize gestures of other categories (e.g., iconic gestures). Besides, we are interested in ameliorating the emotional content of the synthesized speech so as to make the generated speech more persuasive and more natural.

REFERENCES

[1] A. Aly and A. Tapus, "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, 2013.

[2] P. Ekman, *About brows: Emotional and conversational signal*, pp. 169–248. Cambridge, UK: Cambridge University Press, 1979.

[3] R. Cowie and R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech Communication*, vol. 40, pp. 5–32, 2003.

[4] J. Bachorowski, "Vocal expression and perception of emotion," *Current Directions in Psychological Science*, vol. 8, no. 2, pp. 53–57, 1999.

[5] M. Pell and S. Kotz, "On the time course of vocal emotion recognition," *PLoS ONE*, vol. 6, no. 11, 2011.

[6] A. Aly and A. Tapus, "Towards an online fuzzy modeling for human internal states detection," in *Proceedings of the 12th IEEE International Conference on Control, Automation, Robotics, and Vision (ICARCV)*, (Guangzhou, China), 2012.

[7] A. Aly and A. Tapus, *An online fuzzy-based approach for human emotions detection: An overview on the human cognitive model of understanding and generating multimodal actions*, vol. 106. Switzerland: In Intelligent Assistive Robots, Series: Springer Tracts on Advanced Robotics (STAR), S. Mohammed et al. (Eds.), 2015.

[8] I. Murray and J. Arnott, "Implementation and testing of a system for producing emotion-by-rule in synthetic speech," *Speech Communication*, vol. 16, no. 4, pp. 369–390, 1995.

[9] M. Edgington, "Investigating the limitations of concatenative synthesis," in *Proceedings of Eurospeech*, (Greece), 1997.

[10] A. Iida and N. Campbell, "Speech database design for a concatenative text-to-speech synthesis system for individuals with communication disorders," *Speech Technology*, vol. 6, no. 4, pp. 379–392, 2003.

[11] M. Schroder and J. Trouvain, "The German text-to-speech synthesis system Mary: A tool for research, development, and teaching," *Speech Technology*, vol. 6, no. 4, pp. 365–377, 2003.

[12] A. Kendon, *Gesticulation and speech: Two aspects of the process of utterance*, pp. 207–227. 1980.

[13] D. McNeill, *Hand and mind: What gestures reveal about thought*. IL, USA: University of Chicago Press, 1992.

[14] P. Ekman and W. Friesen, "The repertoire of nonverbal behavior: Categories, origins, usage, and coding," *Semiotica*, vol. 1, pp. 49–98, 1969.

[15] A. Kendon, *The study of gesture: Some remarks on its history*, pp. 153–164. NY, USA: Springer, 1983.

[16] C. Pelachaud, "Multimodal expressive embodied conversational agents," in *Proceedings of the 13th Annual ACM International Conference on Multimedia*, (NY, USA), pp. 683–689, 2005.

[17] S. Kopp and I. Wachsmuth, "Synthesizing multimodal utterances for conversational agents," *Computer Animation and Virtual Worlds*, vol. 15, no. 1, pp. 39–52, 2004.

[18] Q. Le, J. Huang, and C. Pelachaud, "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), 2012.

[19] J. Cassell, H. Vilhjálmsson, and T. Bickmore, "BEAT: The behavior expression animation toolkit," in *Proceedings of the SIGGRAPH*, pp. 477–486, 2001.

[20] A. Aly and A. Tapus, "A model for synthesizing a combined verbal and nonverbal behavior based on personality traits in human-robot interaction," in *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, (Tokyo, Japan), pp. 325–332, 2013.

[21] A. Aly and A. Tapus, "Towards an intelligent system for generating an adapted verbal and nonverbal combined behavior in human-robot interaction," *Autonomous Robots*, vol. 39, no. 1, 2015.

[22] J. Cassell, T. Bickmore, L. Campbell, H. Vilhjálmsson, and H. Yan, *Human conversation as a system framework: Designing embodied conversational agents*, pp. 29–63. MA, USA: MIT Press, 2000.

[23] C. Busso and S. Narayanan, "Interrelation between speech and facial gestures in emotional utterances: A single subject study," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2331–2347, 2007.

[24] A. Aly, *Towards an interactive human-robot relationship: Developing a customized robot's behavior to human's profile*. PhD thesis, ENSTA ParisTech, France, 2014.

[25] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "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, 2004.

[26] L. D. Silva, T. Miyasato, and R. Nakatsu, "Facial emotion recognition using multimodal information," in *Proceedings of IEEE International Conference on Information, Communications, and Signal Processing (ICICS)*, vol. 1, (Singapore), pp. 397–401, 1997.

[27] L. Chen, T. Huang, T. Miyasato, and R. Nakatsu, "Multimodal human emotion/expression recognition," in *Proceedings of the 3rd IEEE International Conference on Automatic Face and Gesture Recognition*, (Nara, Japan), pp. 366–371, 1998.

[28] F. Parke, "Computer generated animation of faces," in *Proceedings of the ACM Annual Conference*, vol. 1, (NY, USA), pp. 451–457, 1972.

[29] S. Platt and N. Badler, "Animating facial expressions," *Computer Graphics*, vol. 15, pp. 245–252, 1981.

[30] J. Spencer-Smith, H. Wild, A. Innes-Ker, J. Townsend, C. Duffy, C. Edwards, K. Ervin, N. Merritt, and J. Paik, "Making faces: Creating three-dimensional parameterized models of facial expression," *Behavior Research Methods, Instruments, and Computers*, vol. 33, no. 2, pp. 115–123, 2001.

[31] C. Breazeal, "Towards sociable robots," *Robotics and Autonomous Systems*, vol. 42, pp. 167–175, 2003.

[32] A. Breemen, X. Yan, and B. Meerbeek, "iCat: An animated user-interface robot with personality," in *Proceedings of the 4th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, (Utrecht, Netherlands), 2005.

[33] R. Beira, M. Lopes, M. Praga, J. Santos-Victor, A. Bernardino, G. Metta, F. Becchi, and R. Saltaren, "Design of the robot-cub (iCub) head," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, (USA), pp. 94–100, 2006.

[34] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," in *Proceedings of the IEEE*, vol. 77, pp. 257–286, 1989.

[35] A. Aly and A. Tapus, "An integrated model of speech to arm gestures mapping in human-robot interaction," in *Proceedings of the 14th IFAC Symposium on Information Control Problems in Manufacturing (INCOM)*, (Bucharest, Romania), 2012.

[36] P. Taylor and A. Isard, "SSML: A speech synthesis markup language," *Speech Communication*, vol. 21, pp. 123–133, 1997.

[37] P. Ekman and W. Friesen, *Facial action coding system: A technique for the measurement of facial movement*. CA, USA: Consulting Psychologists Press, 1978.

[38] D. Shichuan, T. Yong, and A. Martinez, "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, 2014.

[39] J. Hewig, D. Hagemann, J. Seifert, M. Gollwitzer, E. Naumann, and D. Bartussek, "A revised film set for the induction of basic emotions," *Cognition and Emotion*, vol. 19, no. 7, pp. 1095–1109, 2005.

[40] J. Coan and J. Allen, *The handbook of emotion elicitation and assessment (Series in affective science)*. NY, USA: Oxford University Press, 2007.

[41] D. Roberts, H. Narayanan, and C. Isbell, "Learning to influence emotional responses for interactive storytelling," in *Proceedings of the AAAI Spring Symposium on Intelligent Narrative Technologies*, (Stanford University, USA), pp. 95–102, 2009.

[42] M. Siegel, C. Breazeal, and M. Norton, "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, 2009.