Biologically Inspired Cognitive Architectures 24 (2018) 107–114

Contents lists available at ScienceDirect

Biologically Inspired Cognitive Architectures

journal homepage: www.elsevier.com/locate/bica

An embodied virtual agent platform for emotional Stroop effect experiments: A proof of concept

A. Oker[a,*], N. Glas[a], F. Pecune[b], C. Pelachaud[c]

[a] Institut Mines Télécom, ParisTech, Paris, France
[b] ArticuLab, Carnegie Mellon University, Pittsburgh, United States
[c] The Institute for Intelligent Systems and Robotics, UMR7222 CNRS, Pierre and Marie Curie University, Paris, France

ABSTRACT

The need for experimentation of facial expression recognition in a more ecological manner necessitates the use of multimodal, interactive experimental stimuli. At the same time, the prerequisite of reproducibility of results and controlled conditions is still mandatory. An embodied conversational agent (ECA) is a pertinent framework that meets all these requirements. The VIB (Virtual Interactive Behavior) Platform is a SAIBA compliant system which supports the real-time generation of multimodal behavior for interacting with socio-emotional virtual agents. We created a new feature for this platform, namely VIB-Ex, which can be used for presenting real-time facial expressions and recording the user's reaction time and interaction while exporting data for statistical purposes. In this paper, we present our proof of concept study in which a 3D male virtual character has been used to convey joyful or sad facial expressions. At the same time, the same character pronounced joyful or sad words in congruence or incongruence with its facial expression in order to trigger an emotional Stroop effect. Only 12 adults were sufficient in order to obtain an emotional Stroop effect within our virtual agent. The results of this study confirmed that the VIB-Ex platform can replicate a robust effect of psychological phenomena concerning recognition of facial expressions. VIB-Ex proves itself to be a suitable and a pertinent tool to perform experiments on a human's automatic process of facial expression recognition. Finally, we discuss the possible future research topics with VIB-Ex to carry out other type of experiments in the field of social cognition.

Introduction

In cognitive sciences, like any other empirical research, reductionism is the main investigation method. Every aspect of human behavior is reduced to its smaller scales and modules in order to investigate its function on the central nervous system. Historically, every cognitive task or process is deliberately dissected from the 'whole' system and used in this manner in experimental paradigms. The same logic applies for the modality of stimuli; although richer modality can be used in cognitive sciences, the major trend remains largely the use of unimodal stimuli in experimentation. From behaviorism of the early 19th century, this approach has the merits of discovering most structural and functional properties of the cognitive systems, such as memory, human language or vision, as well as providing most intriguing insights to some neurological (hemispatial neglect, Alzheimer's…) and psychiatric conditions (schizophrenia, autism…).

However, following Frith's works in the early 90's regarding the theory of mind (ToM; Frith, 1992) and the discovery of mirror neuron system (MNS) by Rizolatti and collaborators (Rizzolatti, Fadiga, Gallese, & Fogassi, 1996), the very unique idea of social cognition as being qualitatively different to any other neurocognitive process has been brought to the daylight. In parallel, the field of emotion recognition research from the facial expressions has been developed (Ekman & Friesen, 1978; Russell, 1994). These three research subjects (ToM, MNS and recognition of facial expressions) have been generally admitted to constitute the core of the social cognition (Van Overwalle, 2009). Thus, the social cognition is what refers "to the mental operations underlying social interactions, which include processes involved in perceiving, interpreting, and generating responses to the intentions, dispositions, and behaviors of others" (Green, Olivier, Crawley, Penn, & Silverstein, 2005).

This is why, following the emergence of social cognition as a whole new field of research, some authors advocated for the need that stimuli used in experiments should be similar to real life as much as possible. In cognitive science, the term "ecological" is used for stimuli which is more multimodal, interactive and similar to real-life experience rather than the unimodal, non-interactive and representational stimuli used in the laboratory environment. As a matter of fact, in 1979 James J.

[*] Corresponding author.
E-mail address: Dr.alioker@gmail.com (A. Oker).

https://doi.org/10.1016/j.bica.2018.04.011
Received 8 December 2017; Received in revised form 25 April 2018; Accepted 26 April 2018

Gibson already used the term "ecological psychology" for the study of how perception is built by prior visual experience and how perception involves the transmission of sensory stimulation to a passive observer to the idea of *perception–action* systems exploring an information-rich environment. Starting with this view of the perceiver as embodied agent, and perceiving as the functioning of perceptual systems, Gibson explores the questions of what is the nature of the environment that is perceived and how are its functionally significant properties detected. Thus, the word "ecological" is commonly used by cognitive science researchers whom use real life-like stimuli in their experiments.

In this paper,

1) We will show that our embodied conversational agent platform the VIB – Virtual Interactive Behavior – (Pecune, Cafaro, Chollet, Philippe, & Pelachaud, 2014) is biologically inspired thanks to its Saiba compliance.
2) We will present a new module namely the "VIB-Ex" in this architecture which can be used for procedural experimentations in social cognition such as emotional Stroop effect.
3) We will present how emotional Stroop effect can be measured in a more ecological way using embodied conversational agents using our platform. The field of facial recognition is one of the most investigated research subjects in social cognition therefore new technologies as the embodied conversational agents can be a real asset in new research avenues. In order to study a facial expression phenomenon, we have chosen emotional Stroop effect because it forces necessarily and automatically participants to recognize facial expressions. Thus, finally we will present our proof of concept which is an empirical study used to see whether our VIB-Ex module can trigger automatic recognition of a virtual agent's facial expressions evaluated by the means of quantitative and inferential statistics.

Our purpose is to provide arguments that an embodied conversational agent (The VIB, Pecune et al., 2014) module we've developed in order to perform behavioral experiments the "VIB-Ex" can provide more naturalistic experimental settings. To do so, we have conducted a validation study for replicating robust psychological phenomena already being brought to the daylight in cognitive sciences: an emotional Stroop effect. Therefore, our primary objective is to conduct a proof of concept of the VIB-Ex module while creating a Stroop effect within an embodied conversational agent. Before presenting the experiment, firstly we will present our platform, then why Stroop effect is a good candidate to evaluate the ability to perceive facial expression automatically and why Stroop effect can be explained by mirror neuron system triggered by a biologically inspired virtual agent architecture.

Facial expression recognition as a complex cognitive process

The ability to understand other's emotion and inner states (the theory of mind) and emotions require recognition of facial expressions undoubtedly. As Frijda (1988) pointed out that while "emotions and feelings are often considered the most idiosyncratic of psychological phenomena…. It can be described in terms of a set of laws" (pg. 349). Following this idea, Ekman (1992) and Izard (1992) propose in accordance with an evolutionary psychology perspective that recognition and expression of an emotion are fundamental in the social behavior of human species and they can be described and characterized. Consequently, information processing demands made by social cognition are different from those made by non-social cognition. According to Adolphs (2001), "Compared to the physical environment in general, the social environment is more complex, less predictable, and, critically, more responsive to one's own behavior." Thus, as a communicative intent or as an expression of inner state, human expressions plays a fundamental and complex role. By the neuroscientific perspective, neural structures involved in social cognition, even particularly in facial expression recognition are multiple and dynamic. For instance, amygdala plays an important role for the appraisal of threat; it is activated when normal subjects view facial expressions of fear but equally in processing the direction of gaze of others (Baron-Cohen et al., 1999) while the insular cortex is involved in disgust and self-consciousness (Phillips & et al., 1997). However, amygdala and the insular cortex does not act alone. Regarding the faces; inferotemporal cortex; particularly superior temporal sulcus and gyrus fusiform area are primarily selective for facial expressions and face detection but also for intended action recognition based on those (Allison, Puce, & McCarthy, 2000). Finally, another region of great interest; the inferior parietal lobule and the caudal sector of the inferior frontal gyrus and the adjacent part of the premotor cortex, has been highlighted for involving neural mechanisms which allow us to directly understand the meaning of the actions and emotions of others by internally replicating ('simulating') them without any explicit reflective mediation (Gallese, Keysers, & Rizzolatti, 2004). According to these authors, when we witness the disgusted facial expressions of someone else, this network helps to activate that part of our insula that is also active when we experience disgust for example. While the cognitive neuroscience of facial expression recognition is not the purpose of our work, it seems clearer that this is a highly demanding cognitive process and researchers in social cognition are demanding the experimental tasks which can help them to investigate these networks involving facial expressions.

In this field, the literature seems to agree that there would be seven fundamental facial expressions: Joy, fear, anger, surprise, sadness, disgust and contempt. Ekman and Friesen (1978) introduced the idea that there would be a prototypical law of features for these fundamental expressions which would be also universal (Ekman, 1994). Of great importance for embodied conversational agents that we present in this paper; Ekman and Friesen (1978) also proposed a taxonomic system of facial muscles named FACS (Facial Action Coding System) which would standardize physical expression of emotions. These codes called 'Action Units' is what is being used to animate most conversational virtual agents today, but not all.

The photography of stimuli provided by Ekman (1993) started mostly to study facial expression recognition impairments of several social cognition disorders in psychiatry. For instance, schizophrenia patients seem to be very impaired to this task (Edwards, Jackson, & Pattison, 2002; Kohler, Walker, Martin, Healey, & Moberg, 2009), as well as autism spectrum disorders (Harms, Martin, & Wallace, 2010) and in natural of pathological aging individuals (Chaby & Narme, 2009). However, most of these studies use unimodal stimulus (e.g. pictures) and lack to investigate ecological nature of facial expression. On this subject, Keysers and Gazzola (2007) noted that "much of the debate in social cognition might result from choosing tasks that isolate the processes of just one route in the laboratory. However, it is essential to start designing tasks that reflect the complexity of social life to test how the social brain forms an integrated whole." Therefore, we believe that Embodied Conversational Agents (ECA) are a pertinent tool for facial recognition research.

The pertinence of the use of ECA's on facial expression experiments

The pertinence of the use of ECA's in social cognition research has been already highlighted (Gratch, 2014). According to Gratch, "virtual humans aspire to simulate the cognitive abilities of people, but also many of the "embodied" aspects of human behavior, more traditionally studied in fields outside of cognitive psychology, such as nonverbal behavior recognition and production".

The need to investigate social cognition in a more ecological manner necessitates the use of multimodal, interactive experimental stimuli (Oker, Courgeon et al., 2015, Oker, Prigent et al., 2015) called naturalistic settings in psychology experiments. In the late 1980s, Ickes and coworkers already suggested the use of the term naturalistic social cognition for studies in which the experimental conditions were socially complex situations (Ickes & Tooke, 1988). The use of 'naturatistic' word is also advocated by Zaki and Ochsner (2009) which implies experimental settings in which two interacting agents (human or machine) can transfer information between them as it would happen in real life. Moreover, in order to create a richer realism in the laboratory, naturalistic settings need more ecological stimuli. As we explained before, by more ecological, we mean richer multimodal stimuli in terms of realism and movement; for instance, animated facial expressions not played by actor but by a system with an established taxonomy based on real human facial muscles. This is because, the prerequisite of reproducibility of results and controlled conditions are still mandatory for any psychological study. Embodied conversational agents (Cassell, 2001) are the most pertinent framework that meets all the requirements: They can simulate human intelligence, communicate by expressing affects, react according to user's actions, and contextualize a human interaction as a whole. They can communicate via verbal and nonverbal behaviors (gaze, facial expression, gesture). Therefore, ECAs can be used in facial expression experiments, as well as empathy or intentionality research (Grynszpan et al., 2011; Jackson, Michon, Geslin, Carignan, & Beaudoin, 2015, for a review see Paiva, Leite, Boukricha, & Wachsmuth, 2017). Moreover, not only stimuli itself are more ecological; the way that the conversation can occur in a continuum makes it more naturalistic. An ECA also makes it possible to control experimental conditions and to observe scientific reproducibility because the fact that ECA programming can also pursue a hierarchical arborescence or adaptive algorithms to user's input if necessary.

Several ECA's has been already successfully deployed in order to provoke and observe psychological responses. For instance, Creed and Beale (2007) investigated the psychological impact of simulated emotional expressions on computer users. Of great importance for the current study, in this work, the authors used the different emotional facial expressions, by making use of Ekman's (2002) work on expression and his taxonomy. However, contrary to our implicit measure (reaction times), they preferred to use explicit and conscious answers using seven-point semantic differential scales. This comparison on implicit versus explicit measures with the presence of humanoid virtual agents is a real research question. For instance, Bailenson et al. (2004) preferred to use both self-report and behavioral measures in order to evaluate the embodied virtual agent's social presence. They've concluded that hat nonverbal behavior may be a more sensitive measure of the copresence and general influence of embodied agents than self-report measures. Unfortunately, this study has not aimed to evaluate more involved nonverbal behaviors, such as facial expressions like ours.

The VIB -Virtual Interactive Behavior-

The VIB -Virtual Interactive Behavior- (Pecune et al., 2014) that we've used in this study is an extensible ECA platform to implement scientific research on complex human interactions thanks to its "Situation, Agent, Intention, Behavior and Animation" (SAIBA) compliance.[1] Therefore, the VIB supports the creation of socio-emotional ECAs. SAIBA is a multimodal behavior generation framework for embodied conversational agents. The framework specifies multimodal generation of embodied virtual agents at a macro-scale, consisting of processing stages on three different levels: (1) planning of a communicative intent, (2) planning of a multimodal realization of this intent, and (3) realization of the planned behaviors. It can accommodate any external module thanks to middlewares like ActiveMQ, Thrift, RabbitMQ communications protocols.

The main contribution of the compliance of the architecture is equally inspired by human cognition: For instance, appraisal theories (for instance the OCC; Ortony, Clore, & Collins, 1990 or Scherer's theory, 2001) can be implemented in the modules called Behavior Planner and Realizers. Behavior Planner describes communicative and expressive intent without any reference to physical behavior by the means of the Function Markup Language (FML). Behavior realizer describes multimodal behaviors as they are to be realized by the final stage of the generation process by the means of the Behavior Markup Language (BML) for this purpose. It provides a general, player-independent description of multimodal behavior that can be used to control our embodied agent. Within this unifying structure, the VIB is defined by a particular set of modules which can give the opportunity to create different architectures in order to animate a virtual agent. These architectures, particularly the content of Behavior planner can be data-driven or procedural as depicted with the examples in Fig. 1.

The VIB platform allows the agent to be an active interactant (speaker or listener). The agents in the platform can display diverse intentions (Callejas, Ravenet, Ochs, & Pelachaud, 2014) through facial expressions (Ochs & Pelachaud, 2013) and gestures (Mancini & Pelachaud, 2008). Not only procedural (or following a linear script) VIB platform can also use machine learning techniques. Machine learning have been implemented in order to drive the multimodal behaviors of the agent during laughing and delivering an emotional speech (Ding, Pelachaud, & Artires, 2013; Ding, Prepin, Huang, Pelachaud, & Artires, 2014). The idea behind is to give the ability to VIB to adapt its behavior according to social constructs (attitude, relationship) that are consistently updated depending on its interlocutor and the dynamic of the interaction. The attitude can be modified on the scale of big-five (a model which suggest that personality can be described by five traits: openness to experience, conscientiousness, extraversion, agreeableness, neuroticism, Costa & McCrae, 1985, 1992) or friendliness (Argyle, 1988). This approach has already been successfully manipulated with an ECA (Cafaro et al., 2012). The VIB platform can also adapt socially to the user and perform dyadic interaction and emerge interpersonal phenomena.

Furthermore, our ECA agent has the capacity to show its engagement during the interaction by choosing its conversation topics (Glas, Prepin, & Pelachaud, 2015) and by making use of hetero-repetition, on expanding the expressivity model by analyzing a large database of motion capture data of expressive multimodal behaviors (Fourati & Pelachaud, 2014), and on modeling group behavior during conversations. Finally, Network communications modules have been developed to interface with external software or to exchange events between different instances of the platform over several software. Currently, different APIs are used such as ActiveMQ and Thrift. For example, the SSI (Social Signal Interpretation Framework, Wagner et al., 2013) can be implemented in order to record, analyze and recognize human behavior in real time, such as gestures, mimics, head nods, and emotional speech, then by triggering appropriate action with artificial intelligence modules like FATIMA (Dias, Mascarenhas, & Paiva, 2014).

Objective of our proof of concept for Vib-Ex module for triggering emotional Stroop

We've tested if the VIB platform can carry out facial expression experiments by replicating robust psychological phenomenon called emotional Stroop and provide arguments in favor of the VIB-Ex's more ecological nature in facial expression recognition experiments. In fact, the emotional Stroop effect is a good candidate for a proof of concept because the fact that it implies automatic and non-deliberate recognition of facial expressions. Event-related potential research in electroencephalography (EEG) shows that the N170 wave is modulated by face (George, Evans, Fiori, Davidoff, & Renault, 1996) and emotional expressions (Kropotov, 2016). The N170 is an ERP component characterized by a negative deflection in wave amplitude that occurs around 170 ms after the presentation of a face. As such, it is a very early marker of automatic, non-deliberate and non-conscious proof of neural

[1] Lead by: Hannes Vilhjalmsson, Norman Badler, Lewis Johnson, Stefan Kopp, Brigitte Krenn, Stacy Marsella, Andrew N. Marshall, Catherine Pelachaud, Hannes Pirker, Kristinn R. Thorisson. See http://www.mindmakers.org/projects/saiba/wiki.

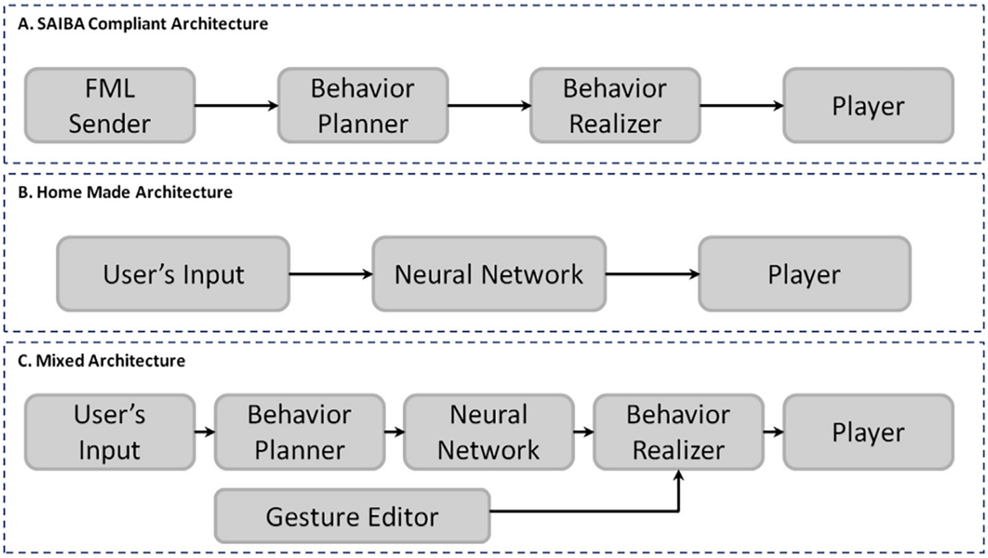


Fig. 1. Examples of architectures created in the VIB (Pecune et al., 2014).

activation when the stimulus is presented to the cognitive system.

In 1935, Stroop has found out that naming a color of a word is delayed if the word itself is another color (for instance the word yellow is written in red and the task is not to read the word ("yellow") but to name the color ("red")). This is one of the most important paradigms in cognitive psychology regarding the study of attention process. The main purpose is to study interference between cognitive process and control/inhibitory capacities. Others have used a mix of word-pictures with congruent and non-congruent conditions semantically related or not (Guttentag & Haith, 1978). The results are striking; it is nearly impossible to comply with the task or not to be delayed in a non-congruent condition.

Emotional Stroop (Watts, McKenna, Sharrock, & Trezise, 1986) is a task in which the same idea of Stroop is maintained but the interference is made by the emotionally significant words while color naming. McKenna and Sharma (1995) has found out also that emotionally significant words disrupt color naming more than neutral words. Other work has dropped out color naming but has used emotionally significant pictures associated with congruent and non-congruent words in priming paradigm (Bower, 1994). Again, in social neuroscience, the emotional Stroop effect has been successfully used in order to study cognitive control mechanism in brain (Ovaysikia, Tahir, Chan, & DeSouza, 2010). However, the modality of the used social stimuli was again unimodal (pictures in which a word is written over the face) by laboratory members.

In this study presented, we used words not written on the screen but pronounced by virtual character expressing joy or sadness in congruence or not with the meaning. Hence, we used the lowest necessary conditions for binary responses to replicate a Stroop effect with more ecological stimuli then cited above.

Automatic emotion recognition and embodied cognition

We also believe that emotional Stroop effect can be discussed in the field of embodied cognition. As a matter of a fact; it can be proposed that given our knowledge of embodied or situated cognition (Varela, Thompson, & Rosch, 1991), in non-pathological subjects, the recognition of facial expression is automatic as being one of the most important expertise from developmental stages. According to Barsalou of embodied cognition theory assumes that perception cannot be explained apart from the action that accompanies it (Barsalou, 2008). In our Stroop experiment carried out with an embodied conversational agent, categorization of the pronounced words as being joyful or sad is the action required (providing dependent variable), while looking at the congruent or non-congruent facial expressions (perception) to these words is the independent variable. Normally, a cognitive system does not require to process facial expression in order to categorize words as being joyful or sad. However, even though the task does not require to evaluate facial expression, because the fact that recognizing a facial expression is automatic, word categorization is forcefully influenced by the nature of this facial expression, as it should be in any emotional Stroop effect.

This automatic and nonconscious recognition activity brings also all the possibilities of what can be said with a joyful and sad face according the knowledge accumulated in this situation or social environment. In an emotional Stroop effect, reading or categorizing words, in our opinion, is a feature of the environment created during the experiment while automatic facial expression recognition is an ability of the organism acquired prior to the experiment. Therefore, in the case of a joyful face pronouncing 'gift' for example, the word would be categorized faster caused by the semantic organization of the joyful face as it is in congruence within the word 'gift'. However, if the joyful face pronounces 'coffin', that would cease semantic organization by being invalid between the sense of expression and the word which would be measured by longer reaction times in our paradigm. Whether, this incongruence is intentional or not (for example, one might want to be ironic or ambiguous), this would always trigger longer reaction times. We will further discuss our results within this framework after the presentation of our proof of concept.

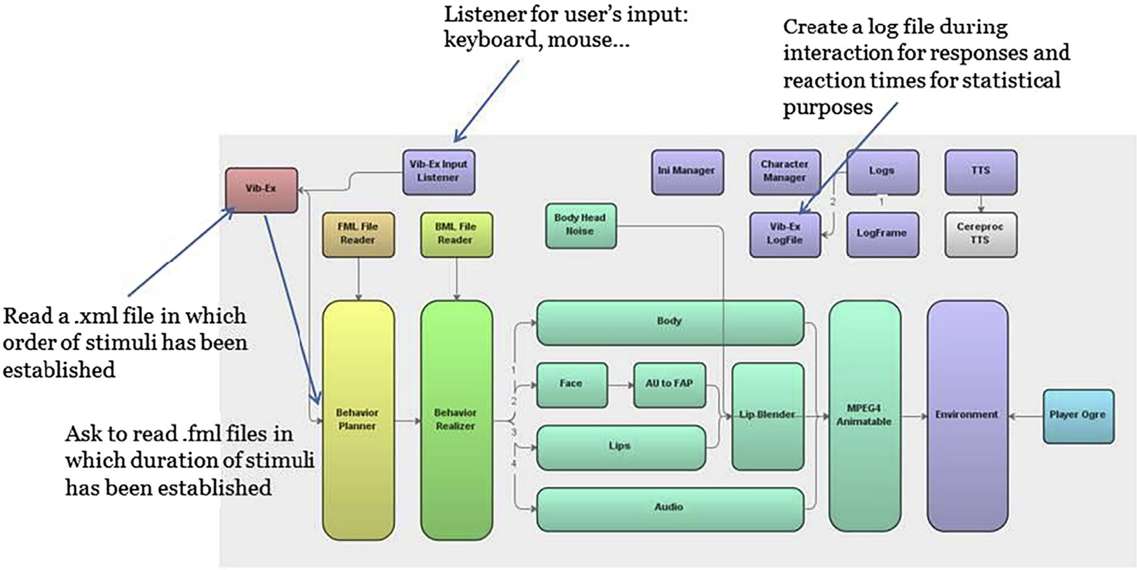
Experiment

Subjects

12 adult subjects (3 female 9 male) have been tested (*mean age* = *30.6*, *standard deviation* = *5.9*) between 18 and 44 years old. They've declared that they have not been suffering from any psychiatric disorder or from any neurocognitive disease at the time of the experimentation. They all had a normal or corrected vision.

Methods

The VIB-Ex module running in the VIB platform has been used in this validation study. The VIB-Ex tested here is based on Disco (Rich &


Listener for user's input: keyboard, mouse...

Create a log file during interaction for responses and reaction times for statistical purposes

Read a .xml file in which order of stimuli has been established

Ask to read .fml files in which duration of stimuli has been established

Fig. 2. Organization and tasks of the Vib-Ex module.

Table 1
The frequency of occurrence and the intensity of words used in our experiment.

	Intensity of valence	Occurrence frequency in French
Joyful (positive) words	4.07 (of 5) (1.34 std.dev)	33.14 (of million occurrence)
Sad (negative) words	4.21 (of 5) (1.26 std.dev)	35.6 (of million occurrence)

Sidner, 2012). Disco is a dialog manager combining hierarchical task networks with dialog trees which improves the authoring process. It works with an external .xml file to be called in system. We have added functionalities to the VIB's implementation of Disco in order to proceed this procedural experience. Our version now allows easier authoring experimental conditions and working on the duration of the stimuli (verbal and nonverbal). It also allows calculating the reaction times of user's input (and the duration of input as well) and export the results in a .csv file. The main structure of procedural experiment would be determined in VIB-Ex (see Fig. 2). These modifications have been made within the dialog management module (Glas et al.'s (2015) work in which fixed task structure of Disco was overwritten and thus, topic management is controlled by the external topic selection module. Thereby, these modifications altogether open possibility of experimentations in psychology adding flexibility and resulting in a more adaptive and dynamic interaction with the virtual agent.

Stimuli

Words pronounced by the virtual agent

32 items have been chosen from a database based on a study of emotional and intensity evaluation of most frequently used words in French. These words have been evaluated by a population of more than 100 native French speakers (corpus Giraud, 2011). Half of them have been evaluated as being sad and negative while the other half have been evaluated as being joyful and positive. As it can be observed in the organization of the VIB system (graphic 1), the lip blender operates with the text to speech module to carry on a morphing of movements for the words being pronounced according the appropriate dynamics. Thus, the labial movements are rendered over two fronts: the facial expression (joy or sadness) and the word (the duration of the words and upper and lower muscle implications according the vowel).

The pronunciation of the words was carried out by a non-prosodic speech synthesis (Cereproc TTS- text to speech). So, there was no vocal tone or any prosodic information which have been coded in the Cereproc system (https://www.cereproc.com/). The speech rate was kept at the default setting for all participants. Naturally, the order of words was selected pseudo randomly across participants. The frequency (of occurrence) and the intensity of these words is the following:

Emotions expressed by the virtual agent

The behavior realizer module of the VIB platform has been used to render dynamic facial expressions of emotions. The resulting facial expressions were modeled as a set of action units based on the Facial Action Coding System – FACS (Ekman, 2002; Ekman & Friesen, 1978). As presented earlier FACS is a model that describes and classifies characteristics of human facial expressions. They are characterized by the activation of Action Units (AUs), each of them corresponding to the contraction of one or several facial muscles. In our experiment, the AU muscle correspondences are the following (Table 2):

Concerning morphological information displayed, there are several parameters implemented in the VIB system: amplitude of contractions, cheek raising, duration of the expression, mouth opening, symmetry of the lip corner, lip press and the velocity of the onset and offset of the expression. For instance, these parameters allow one to differentiate an amused smile (also called felt, Duchenne, enjoyment, or genuine smile), a polite smile (also called non-Duchenne, false, social, masking, or controlled smile) and an embarrassed smile. These expressions have been subject to user perception testing and were validated (Ochs & Pelachaud, 2013 for smile and Schröder et al., 2011 for sadness).

Table 2
The taxonomic description of action units and facial muscles used in our experiment for joy and sadness.

	Action Units according to FACS	Muscles involved
Joy	AU6 caused cheek raiser AU25 caused nasolabial deepener AU12 caused lip corner puller	*orbicularis oculi, pars orbitalis* contraction *depressor labii inferioris* *zygomaticus major* contraction
Sadness	AU1 caused inner brow raiser AU4 caused brow lowerer AU15 caused lip corner depressor	*Frontalis, pars medialis* contraction *corrugator supercilii, depressor supercilii* *depressor anguli* contraction

Procedure and task

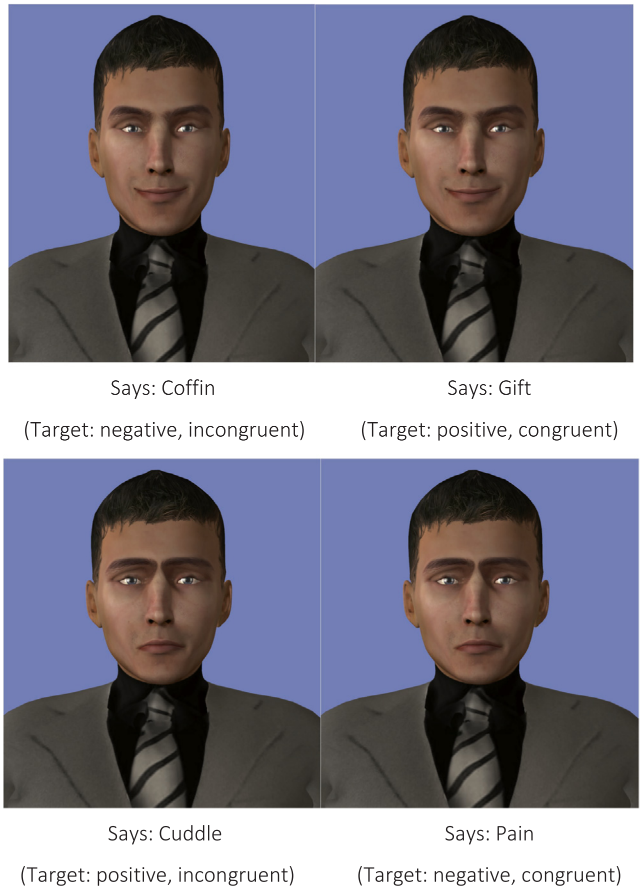
The experiment was carried out on a Dell microcomputer (Intel Xeon® CPU @ 2.40 GHz (2 processors) with 10 Gb RAM and NVIDIA GeForce GTX 560 Ti @ 822 MHz with 4 Gb RAM) with a 24″ monitor on 1920 × 1080 screen resolution displayed with a refresh rate of 60 Hz. The monitor used was not horizontal but vertical as our andromorphous virtual agent. We've also stabilized the chair so that subjects could stay still and remain at a same distance between them and the monitor (approx. 60 cm). This setting provides approximately a visual angle of 48 degrees for width and 28 degrees of height for overall screen.

We've provided a consent form before every subject proceeded to testing. It has been explained to the subjects that the virtual character named Rodrigue would explain the task itself. The virtual agent has started the experimentation by thanking the participants and explaining what they have to do: "You have to look at my face, listen to what I'm saying and indicate by using the keyboard whether the word I pronounce is joyful or sad".

The experimental session has consisted of 36 game turns (trials) following 4 training turns that are excluded from further analyses. The presentation of conditions as well as the order of the nature of words were randomized for every subject.

Experimental conditions

In this study, we have two independent variables within two modalities: Congruence between the word and the expression or incongruence between the word and the expression and the valence of the words, joyful/positive and sad/negative. These nested modalities are represented in the following Fig. 3.



Says: Coffin

(Target: negative, incongruent)

Says: Gift

(Target: positive, congruent)

Says: Cuddle

(Target: positive, incongruent)

Says: Pain

(Target: negative, congruent)

Fig. 3. Example of our stimuli used in the experiment.

Results

Mean correct response latencies and error rates were calculated across the subjects for each experimental condition. Latencies more than three standard deviations above or below the mean were excluded (less than 5% of the data) per subject. The mean correct latencies and error rates for the different experimental conditions are presented in Table 1.

Separate repeated measures analyses of variance were performed on latencies and error rates with Word type (joyful/positive versus sad/negative) and Congruence type (congruent versus incongruent) as the within-subject variable.

The analysis of the error rates revealed no main effect or interaction **$F < 1$**. This result could be explained by ceiling effects since the overall level of correct responses was near 98%. This is understandable given that Stroop effect is only observed within reaction times.

The analysis of the latencies revealed a significant main effect of Congruence Type, **$F(1, 11) = 14.8$, $p < 0.005$, η2G (generalized eta square) $= 0.27$** but no interaction with Word type, which means that Stroop effect is independent from of the target word's valence. This result indicates that there is a significant Stroop effect between incongruent condition (**2352 ms**) and congruent condition (**2175.5 ms**). The fact that we found a Stroop effect with 12 subjects is highly encouraging but not surprising. The Stroop effect is a very robust phenomenon known since 1935, it can be found with a very similar sample size. As the Stroop effect is an already known phenomenon, our scope in this proof of concept is to highlight the fact that the embodied conversational agents can carry out facial expression recognition experiments in a more ecological way and the Stroop effect can be found with our system. Moreover, the analysis of variance is formal, and the effect size is within the limits of the general use in the literature (Weinberg & Abramowitz, 2002) (Table 3).

Discussion

Our purpose in this study has been to emphasize that embodied conversational agents form one of the most pertinent tools for facial expression recognition experimentations. To support this claim, we've conducted a validation study which consists of replicating a most common psychological phenomenon: emotional Stroop. We've chosen to see whether emotional Stroop can be replicated and measured with a SAIBA compliant embodied conversational agent platform. In order to do this, we've conducted an implementation and modification of the dialog manager module (Disco module (Glas et al., 2015)), we've also added to VIB system the capacity of carrying out behavioral experimentations by authoring experimental conditions, manipulating the timing of social stimuli presentation, creating a logging system which is compatible with statistical analysis. To our knowledge, this is the first time a SAIBA compliant embodied conversation agent system (VIB) is bundled with an experimentation module (VIB-Ex) which can carry out psychological experimentations with success. In our validation study with predetermined procedures, we have found out that when a virtual agent pronounces words whose emotional valence is not congruent with

Table 3
Mean response times (RTs) and error rates (ERs) for each experimental condition with standard errors.

Condition	Error Rate%	Error Rate (Std. Dev)	Reaction Times (Mean ms)	Average (ms) Congruence/ Incongruence	Reaction Times (Mean Std.Dev)
Congruence/pos	%2	%2	2150	2175.5	206
Congruence/neg	%1	%1	2201		195
Incongruence/pos	%1	%1	2393	2352	337
Incongruence/neg	%3	%3	2311		362

the agent's facial expression the response times are slower (mean of 2352 ms) than when it is congruent, as a typically Stroop task would work (mean of 2175.5 ms). This effect is so robust in cognitive psychology, that we were able to replicate it with 12 subjects with a moderate to important effect size (generalised eta squared 0.27), primarily because of our system's render of facial expressions is accurate (Ochs & Pelachaud, 2013 for the joyful expression; Schröder et al., 2011 for sadness).

As it is pointed out by Frijda's views, the "action readiness" (1986) seems to be compatible with embodied cognition theory. The detection of a facial expression even not relevant to what the individual should do in a task gives rise to an embodied readiness for action. In our case, it can be proposed that non-pathological cognitive system is *craving* for social cues and automatically organizing the system into what would be adequate afterwards. In our validation study, the non-congruence condition would cease this organization and provides longer reaction times. While every modality is perceived correctly (our paradigm provoked very few errors), because non-pathological subjects have an expertise to recognize emotions automatically and to understand the meaning of words; a non-congruent experimental condition would also cease affordance organization of cognitive system tested by the virtual agent used in the experiment.

Here, in our paradigm, there is a historical theoretical and biological construct which can account for the features of the 'social' environment presented to the subjects: on one hand a multimodal affective virtual agent displaying facial muscle contractions that are presumed to be universal, and on the other hand the subjects are automatically giving a meaning (joy/sadness) to these contractions because of the evolutionary result of nervous system's organization in a social environment.

Neurocognitive architecture behind emotional Stroop effect made by Vib-Ex

According to Gratch (2014), most recent research on embodied cognition "emphasizes that even the state of our bodies contains important information that informs and influences cognitive process (Wilson, 2002) and this work has begun to influence computational models of cognitive processes (e.g., see Ritter & Young, 2001)" as well. Situated or embodied view of cognition (Barsalou, 2008) accounts also for social cognition. According to Barsalou (2008), "when a facial expression or posture is adopted it elicits associated mental states" (pg. 630). Social interaction seems to be also underlying by mirror neuron system (MNS). It seems that during observation of an action (in our experiment facial expression) the input is compared to one's own motor repertoire, and when they match, the action's goal is identified (Rizzolatti, Fogassi, & Gallese, 2001; Van Overwalle & Baetens, 2009). On the peculiar subject of other's emotion recognition, it has also been proposed that a mirror neuron system can account for understanding facial expressions (Van der Gaag, Minderaa, & Keysers, 2007)[2].

This is why, an increasing number of researchers in psychiatry is interested in embodied conversational agents (Brunet-Gouet, Oker, Martin, Grynszpan, & Jackson, 2016); and/or social robots (Raffard et al., 2016). It is well known that schizophrenia patients have impairments for recognizing facial expressions (Edwards et al., 2002) and empathetic intentions of others (Berrada-Baby et al., 2016). Adding to that, it is largely acknowledged that autistic persons have heterogeneous difficulties to perform imitations (Nadel, 2002). In the field of neuroscience, virtual agents' display of VIB platform has already been used as stimuli (Aranyi, Pecune, Charles, Pelachaud, & Cavazza, 2016) in order to explore a neurofeedback remediation alternative.

Future research

Hence, we believe that the systems like the VIB platform can be a gold standard on naturalistic artificial emotional intelligence in order to carry out not only experimentation on facial expressions but social cognition experimentations altogether. Although this is not the case in our study, artificial emotional intelligence can be part of VIB-Ex. As VIB is a Saiba compliant platform, any artificial intelligence module like FATIMA (Dias et al., 2014) can be integrated with great use. For instance, Fatima endows agents with planning capabilities that can use emotions to act upon user's behaviors. In other words, the agent's affective state can be updated during the interaction according to user's emotional state and can generate coherent or incoherent output if necessary.

Scientific reductionism that we mentioned earlier is not the only way for investigating complex social cognition like human-human or human-computer interaction. Evaluating separated constructs can be an important asset in social cognitive research; however, it cannot evaluate complex interactive situations that include naturalistic stimuli (like facial expressions with motion, prosody, discourse, body movements and the contextualization of those altogether). According to Schilbach et al. (2013), a second-person approach is needed to be adopted in psychological research regarding interaction and emotional engagements between people. Embodied conversational agents like the VIB platform can pursue further psychological research by permitting to emerge interpersonal phenomena between two persons.

While our study was a procedural case study to underline the important naturalistic nature of the VIB platform and to verify that our VIB-Ex module can carry out experiments for facial expression recognition, our future research will concentrate on how VIB-ex can be programmed in order to be used in social cognition studies, as theory of mind, empathy and attribution bias in healthy or pathological populations.

Acknowledgment

This work was partially performed within the Labex SMART (ANR-11-LABX-65) supported by French state funds managed by the ANR within the Investissements d'Avenir programme under reference ANR-11-IDEX-0004-02. It has also been partially funded by the French National Research Agency project MaClasse 3.0.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.bica.2018.04.011.

References

Adolphs, R. (2001). The neurobiology of social cognition. *Current Opinion in Neurobiology, 11*(2), 231–239.

Allison, T., Puce, A., & McCarthy, G. (2000). Social perception from visual cues: role of the STS region. *Trends in Cognitive Sciences, 4*, 267–278.

Aranyi, G., Pecune, F., Charles, F., Pelachaud, C., & Cavazza, M. (2016). Affective interaction with a virtual character through an fNIRS brain-computer interface. *Frontiers in Computational Neuroscience, 10*.

Argyle, M. (1988). *Bodily communication.* New York: Methuen and Company.

Bailenson, J. N., Aharoni, E., Beall, A. C., Guadagno, R. E., Dimov, A., & Blascovich, J. (2004, October). Comparing behavioral and self-report measures of embodied agents' social presence in immersive virtual environments. In *Proceedings of the 7th Annual International Workshop on PRESENCE* (pp. 1864–1105).

Baron-Cohen, S., Ring, H. A., Wheelwright, S., Bullmore, E. T., Brammer, M. J., Simmons, A., et al. (1999). Social intelligence in the normal and autistic brain: an fMRI study. *European Journal of Neuroscience, 11*(6), 1891–1898.

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology, 59*, 617–645.

Berrada-Baby, Z., Oker, A., Courgeon, M., Urbach, M., Bazin, N., Amorim, M. A., et al. (2016). Patients with schizophrenia are less prone to interpret virtual others' empathetic questioning as helpful. *Psychiatry Research, 242*, 67–74.

Bower, G. H. (1994). Some relations between emotions and memory. *The Nature of Emotion Fundamental Questions,* 303–305.

Brunet-Gouet, E., Oker, A., Martin, J. C., Grynszpan, O., & Jackson, P. L. (2016). Editorial:

[2] The inferior frontal gyrus (IFG) and posterior parietal cortex (PPC) for motor components of facial expressions being part of mirror neuron system and amygdala and insula for emotional components to be exact (see Van der Gaag et al., 2007).

Advances in virtual agents and affective computing for the understanding and remediation of social cognitive disorders. *Frontiers in Human Neuroscience, 9*.

Cafaro, A., Vilhjálmsson, H. H., Bickmore, T., Heylen, D., Jóhannsdóttir, K. R., & Valgarðsson, G. S. (2012). First impressions: Users' judgments of virtual agents' personality and interpersonal attitude in first encounters. *International conference on intelligent virtual agents* (pp. 67–80). Berlin Heidelberg: Springer.

Callejas, Z., Ravenet, B., Ochs, M., & Pelachaud, C. (2014). A computational model of social attitudes for a virtual recruiter. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems* (pp. 93–100). International Foundation for Autonomous Agents and Multiagent Systems.

Cassell, J. (2001). Embodied conversational agents: Representation and intelligence in user interfaces. *AI Magazine, 22*(4), 67.

Chaby, L., & Narme, P. (2009). Processing facial identity and emotional expression in normal aging and neurodegenerative diseases]. *Psychologie & neuropsychiatrie du vieillissement, 7*(1), 31–42.

Costa, P. T., & McCrae, R. R. (1985). The NEO personality inventory: Manual, form S and form R. *Psychological Assessment Resources*.

Costa, P. T., & McCrae, R. R. (1992). Four ways five factors are basic. *Personality and Individual Differences, 13*(6), 653–665.

Creed, C., & Beale, R. (2007). Psychological responses to simulated displays of mismatched emotional expressions. *Interacting with Computers, 20*(2), 225–239.

Dias, J., Mascarenhas, S., & Paiva, A. (2014). Fatima modular: Towards an agent architecture with a generic appraisal framework. *Emotion modeling* (pp. 44–56). Springer International Publishing.

Ding, Y., Pelachaud, C., Artires, T. (2013). Modeling multimodal behaviors from speech prosody. In *13th International Conference of Intelligent Virtual Agents - IVA*.

Ding, Y., Prepin, K., Huang, J., Pelachaud, C., Artires, T. (2014). Laughter animation synthesis. In *International Conference on Autonomous Agent and Multi- Agent Systems (AAMAS)*.

Edwards, J., Jackson, H. J., & Pattison, P. E. (2002). Emotion recognition via facial expression and affective prosody in schizophrenia: a methodological review. *Clinical Psychology Review, 22*(6), 789–832.

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion, 6*(3–4), 169–200.

Ekman, P. (1993). Facial expression and emotion. *American Psychologist, 48*(4), 384.

Ekman, P. (1994). Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique.

Ekman, P. (2002). *Facial action coding system.* Salt Lake City: A Human Face.

Ekman, P., & Friesen, W. V. (1978). *Manual for the facial action coding system.* Consulting Psychologists Press.

Fourati, N., & Pelachaud, C. (2014). Emilya: Emotional body expression in daily actions database. In *LREC* (pp. 3486–3493).

Frijda, N. H. (1986). *The emotions: Studies in emotion and social interaction.* Paris: Maison de Sciences de l'Homme.

Frijda, N. H. (1988). The laws of emotion. *American Psychologist, 43*(5), 349.

Frith, C. (1992). *The cognitive neuropsychology of schizophrenia.* London: Lawrence Erlbaum Associates.

Gallese, V., Keysers, C., & Rizzolatti, G. (2004). A unifying view of the basis of social cognition. *Trends in Cognitive sciences, 8*(9), 396–403.

George, N., Evans, J., Fiori, N., Davidoff, J., & Renault, B. (1996). Brain events related to normal and moderately scrambled faces. *Cognitive Brain Research, 4*(2), 65–76.

Giraud, A. (2011). Influence de stimuli émotionnels dans l'alternance entre tâches. Master 2 Professionnalisant Psychologie et Neuropsychologie des Perturbations Cognitives. Université d'Aix en Provence).

Glas, N., Prepin, K., & Pelachaud, C. (2015). Engagement driven topic selection for an information-giving agent. In Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2015- goDial).

Gratch, J. (2014). Understanding the mind by simulating the body: Virtual humans as a tool for cognitive science research. In *Oxford handbook of cognitive science*.

Green, M. F., Olivier, B., Crawley, J. N., Penn, D. L., & Silverstein, S. (2005). Social cognition in schizophrenia: Recommendations from the measurement and treatment research to improve cognition in schizophrenia new approaches conference. *Schizophrenia Bulletin, 31*(4), 882–887.

Grynszpan, O., Perbal, S., Pelissolo, A., Fossati, P., Jouvent, R., Dubal, S., et al. (2011). Efficacy and specificity of computer-assisted cognitive remediation in schizophrenia: A meta-analytical study. *Psychological Medicine, 41*(01), 163–173.

Guttentag, R. E., & Haith, M. M. (1978). Automatic processing as a function of age and reading ability. *Child Development,* 707–716.

Harms, M. B., Martin, A., & Wallace, G. L. (2010). Facial emotion recognition in autism spectrum disorders: A review of behavioral and neuroimaging studies. *Neuropsychology Review, 20*(3), 290–322.

Ickes, W., & Tooke, W. (1988). The observational method: Studying the interaction of minds and bodies.

Izard, C. E. (1992). Basic emotions, relations among emotions, and emotion-cognition relations.

Jackson, P. L., Michon, P. E., Geslin, E., Carignan, M., & Beaudoin, D. (2015). EEVEE: The empathy-enhancing virtual evolving environment. *Frontiers in Human Neuroscience, 9*.

Keysers, C., & Gazzola, V. (2007). Integrating simulation and theory of mind: from self to social cognition. *Space, 8*, 108–114.

Kohler, C. G., Walker, J. B., Martin, E. A., Healey, K. M., & Moberg, P. J. (2009). Facial emotion perception in schizophrenia: A meta-analytic review. *Schizophrenia Bulletin* sbn192.

Kropotov, J. D. (2016). *Functional neuromarkers for psychiatry: Applications for diagnosis and treatment.* Academic Press.

Mancini, M., & Pelachaud, C. (2008, May). Distinctiveness in multimodal behaviors. In *Proceedings of 166). International Foundation for Autonomous Agents and Multiagent Systems*.

McKenna, F. P., & Sharma, D. (1995). Intrusive cognitions: An investigation of the emotional Stroop task. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(6), 1595.

Nadel, J. (2002). Imitation and imitation recognition: Functional use in preverbal infants and nonverbal children with autism. In *The imitative mind: Development, evolution, and brain bases*, (pp. 42–62).

Ochs, M., & Pelachaud, C. (2013). Socially aware virtual characters: The social signal of smiles [social sciences]. *IEEE Signal Processing Magazine, 30*(2), 128–132.

Oker, A., Courgeon, M., Prigent, E., Eyharabide, V., Bazin, N., Urbach, M., et al. (2015). A virtual reality study of help recognition and metacognition with an affective agent. *International Journal of Synthetic Emotions (IJSE), 6*(1), 60–73.

Oker, A., Prigent, E., Courgeon, M., Eyharabide, V., Urbach, M., Bazin, N., et al. (2015). How and why affective and reactive virtual agents will bring new insights on social cognitive disorders in schizophrenia? An illustration with a virtual card game paradigm. *Frontiers in Human Neuroscience, 9*.

Ortony, A., Clore, G. L., & Collins, A. (1990). *The cognitive structure of emotions.* Cambridge University Press.

Ovaysikia, S., Tahir, K. A., Chan, J. L., & DeSouza, J. F. (2010). Word wins over face: emotional Stroop effect activates the frontal cortical network. *Frontiers in Human Neuroscience, 4*.

Paiva, A., Leite, I., Boukricha, H., & Wachsmuth, I. (2017). Empathy in virtual agents and robots: A survey. *ACM Transactions on Interactive Intelligent Systems (TiiS), 7*(3), 11.

Pecune, F., Cafaro, A., Chollet, M., Philippe, P., & Pelachaud, C. (2014). Suggestions for extending saiba with the vib platform. In *Workshop on architectures and standards for IVAs, held at the '14th international conference on intelligent virtual agents (IVA)'*, Bielefeld eCollections (pp. 16–20).

Phillips, M. L., et al. (1997). A specific neural substrate for perceiving facial expressions of disgust. *Nature, 389*, 495–498.

Raffard, S., Bortolon, C., Khoramshahi, M., Salesse, R. N., Burca, M., Marin, L., et al. (2016). Humanoid robots versus humans: How is emotional valence of facial expressions recognized by individuals with schizophrenia? An exploratory study. *Schizophrenia Research*.

Rich, C., & Sidner, C. L. (2012). Using collaborative discourse theory to partially automate dialogue tree authoring. *International conference on intelligent virtual agents* (pp. 327–340). Berlin Heidelberg: Springer.

Ritter, F. E., & Young, R. M. (2001). Embodied models as simulated users: Introduction to this special toward a second-person neuroscience. *Behavioral and Brain Sciences, 36*(04), 393–414.

Rizzolatti, G., Fogassi, L., & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews Neuroscience, 2*(9), 661–670.

Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive brain research, 3*(2), 131–141.

Russell, J. A. (1994). Is there universal recognition of emotion from facial expressions? A review of the cross-cultural studies. *Psychological Bulletin, 115*(1), 102.

Scherer, K. R. (2001). Appraisal considered as a process of multilevel sequential checking. *Appraisal Processes in Emotion: Theory, Methods, Research, 92*(120), 57.

Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., & Vogeley, K. (2013). Toward a second-person neuroscience 1. *Behavioral and brain sciences, 36*(4), 393–414.

Schröder, E. Bevacqua, R. Cowie, F. Eyben, H. Gunes, D. Heylen, M. ter Maat, G. McKeown, S. Pammi, M. Pantic, C. Pelachaud, B. Schüller, E. de Sevin, M. Valstar and M. Wollmer, Building Autonomous Sensitive Artificial Listeners, In *IEEE Transactions of Affective Computing, October* (pp. 134–146).

Van der Gaag, C., Minderaa, R. B., & Keysers, C. (2007). Facial expressions: What the mirror neuron system can and cannot tell us. *Social Neuroscience, 2*(3–4), 179–222.

Van Overwalle, F. (2009). Social cognition and the brain: A meta-analysis. *Human Brain Mapping, 30*(3), 829–858.

Van Overwalle, F., & Baetens, K. (2009). Understanding others' actions and goals by mirror and mentalizing systems: a meta-analysis. *Neuroimage, 48*(3), 564–584.

Varela, F. J., Thompson, E., & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience.* Cambridge, MA: MIT Press.

Wagner, J., Lingenfelser, F., Baur, T., Damian, I., Kistler, F., & André, E. (2013, October). The social signal interpretation (SSI) framework: multimodal signal processing and recognition in real-time. In *Proceedings of the 21st ACM international conference on Multimedia* (pp. 831–834). ACM.

Watts, F. N., McKenna, F. P., Sharrock, R., & Trezise, L. (1986). Colour naming of phobia related words. *British Journal of Psychology, 77*(1), 97–108.

Weinberg, S. L., & Abramowitz, S. K. (2002). *Data analysis for the behavioral sciences using SPSS.* Cambridge University Press.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review, 9*, 625–636.

Zaki, J., & Ochsner, K. (2009). The need for a cognitive neuroscience of naturalistic social cognition. *Annals of the New York Academy of Sciences, 1167*(1), 16–30.

Further reading

Gibson, J. J. (2014). *The ecological approach to visual perception: Classic edition.* Psychology Press.