Int J Soc Robot (2012) 4:29–51
DOI 10.1007/s12369-011-0134-7

Visuo-auditory Multimodal Emotional Structure to Improve Human-Robot-Interaction

José Augusto Prado · Carlos Simplício · Nicolás F. Lori · Jorge Dias

Accepted: 19 November 2011 / Published online: 20 December 2011

Abstract We propose an approach to analyze and synthesize a set of human facial and vocal expressions, and then use the classified expressions to decide the robot's response in a human-robot-interaction. During a human-to-human conversation, a person senses the interlocutor's face and voice, perceives her/his emotional expressions, and processes this information in order to decide which response to give. Moreover, observed emotions are taken into account and the response may be aggressive, funny (henceforth meaning humorous) or just neutral according to not only the observed emotions, but also the personality of the person. The purpose of our proposed structure is to endow robots with the capability to model human emotions, and thus several subproblems need to be solved: feature extraction, classification, decision and synthesis. In the proposed approach we integrate two classifiers for emotion recognition from audio and video, and then use a new method for fusion with the social behavior profile. To keep the person engaged in the interaction, after each iterance of analysis, the robot synthesizes human voice with both lips synchronization and facial expressions. The social behavior profile conducts the personality of the robot. The structure and work flow of the synthesis and decision are addressed, and the Bayesian networks are discussed. We also studied how to analyze and synthesize the emotion from the facial expression and vocal expression. A new probabilistic structure that enables a higher level of interaction between a human and a robot is proposed.

Keywords Visual perception · Auditory perception · Emotion recognition · Multimodal interaction · Social behavior profile · Bayesian networks

The authors gratefully acknowledge support from Institute of Systems and Robotics at University of Coimbra (ISR-UC), Portuguese Foundation for Science and Technology (FCT) [SFRH/BD/60954/2009, Ciencia2007, PTDC/SAU-BEB/100147/2008], and Polytechnical Institute of Leiria (IPL).

J.A. Prado (✉) · C. Simplício · J. Dias
Institute of Systems and Robotics, University of Coimbra, Coimbra, Portugal
e-mail: jaugusto@isr.uc.pt

J. Dias
e-mail: jorge@isr.uc.pt

C. Simplício
Institute Polytechnic of Leiria, Leiria, Portugal
e-mail: carlos.simplicio@ipleiria.pt

N.F. Lori
Institute of Biomedical Research in Light and Image (IBILI), Faculty of Medicine, University of Coimbra, Coimbra, Portugal
e-mail: nflori@fmed.uc.pt

1 Introduction

As natural human communication is mostly contactless, contactless interfaces will be used to reduce the estrangement between humans and machines. Definition of human natural communication is beyond spoken communication. In a face to face interaction between humans, several modalities are normally used, for example: body posture, gestures, gaze, vocalization, and facial expressions. Our focus is to improve the interaction between human and machine by exploring the non-verbal cues, namely facial and vocal expressions. Researchers in this field usually use the term "verbal" as meaning "concerned with words", and do not use "verbal communication" as a synonym for oral or spoken communication. Vocal sounds that are not considered to be words, such as a grunt, or singing a wordless note, are nonverbal.

Thus, the analysis of voice without concern to the words is called nonverbal.

In this study we made an extensive analysis of how humans deal with emotions and feelings, developed a robotic system that can address those issues; developed a Bayesian real time classifiers of emotions from audio and video; developed a Bayesian mixture model to fuse both modalities; then constructed a synthesis with an avatar capable of lips synchronization during speech; and finally different behaviors of the robot were explored and the reaction of the users was tested.

In *visual perception* we focused on facial expression recognition while in *auditory perception* we focused on vocal expression recognition. It is known that the *emotion recognition* problem, in both modalities, depends on two subproblems: the sensory processing that does the extraction of features from the input signal, and the classification of these features across a defined scope. After the classification, the robot reacts according to the human recognized emotion and also according to a robotic *social behavior profile* (*SBP*), see Fig. 1. Moreover, the interaction was supported by synthesizing the vocal and facial expressions, and also lips synchronization. For both analysis and synthesis, a Bayesian framework was used. Our contribution goes beyond the development of novel Bayesian networks, since we also used the results of those Bayesian networks to improve the decision process.

In Sect. 1 the problem of emotion recognition, emotion synthesis and robotic behavior were stated, a brief overview of our solution was also presented. In Sect. 2 we present a study of how humans deal with emotions and feelings. Moreover, still in Sect. 2, it is presented a state-of-the-art on human-robot-interaction via visual and auditory channels; and an overview of the probabilistic framework we used is presented at the end of this section. Section 3 describes how the visual analysis is performed, the explanation starts with image sensory processing and reaches the classification of facial expressions. Results of the facial expression classifier are presented in Sect. 3. Section 4 describes how the auditory analysis is performed, the description starts with audio sensory processing and reaches the classification of vocal expressions. Results of the vocal expression classifier are presented in Sect. 4. Later in Sect. 5 the results of both analysis are combined with a third *random variable* that stands for the robot *social behavior profile*. The fusion of the two perceived modalities with the *social behavior profile (SBP)* of the robot is described on Sect. 5.1. After the fusion, the response that the robot will execute is decided. The robot's response will depend on the robot's learning, so it will vary according to the outputs of the classifiers and also according to the given *social behavior profile*. The robot will then perform the decided response according to what is explained on Sect. 5.2. Notice that the way the robot responds will vary, but the context of each response will not vary. This is due to the fact that our purpose in this work is to explore non-verbal capabilities in human-robot interaction. In Sect. 6 it is explained how the learning phase takes place for audio and video. In Sect. 7 our assessments are defined. In Sect. 8 the experiments are presented and results are shown and discussed. Finally in Sect. 9 conclusions are presented and some future work is proposed.

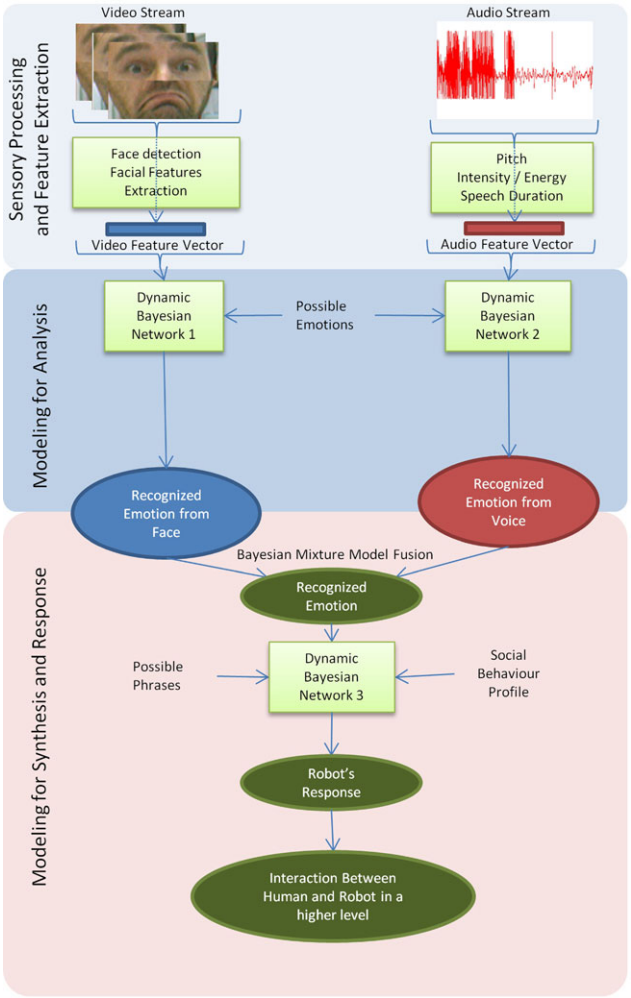

Fig. 1 From audio and video input to robot response, resulting in a higher level of Human-Robot-Interaction. We propose novel Bayesian Models to classify the facial and vocal expressions used during the interaction in real time. For *auditory perception*, a Dynamic Bayesian Network (DBN) is proposed, named DBN2, this network uses information from an audio signal; its outputs are probabilities of vocal expressions. For *visual perception*, DBN1 is proposed where certain local distortions presented over the human face are the evidences to infer the person's facial expression. The information from both networks are combined with *SBP* on the third proposed DBN, DBN3 is responsible for fusion and also decision of the robot response

2 Current State of the Art and Context

2.1 Emotive Robots

There has never been any doubt about the importance of emotions in human behavior, especially in human relationships. The past decade, however, has seen a great deal of progress in developing computational theories of emotion that can be applied to building robots and avatars that interact emotionally with humans. According to the mainstream of such theories [1], emotions are strongly intertwined with other cognitive processing, both as antecedents (emotions affect cognition) and consequences (cognition affects emotions). The robot Autom [2], was designed for extended use in homes as a weight-loss advisor and coach. Autom builds on the research by Bickmore on long-term social interaction and behavior change using avatars. Recently, in [3] was presented the SEMAINE API as a framework for enabling the creation of simple or complex emotion oriented systems. Their framework is rooted in the understanding that the use of standard formats is beneficial for interoperability and reuse of components. They show how system integration and reuse of components can work in practice. An implementation of an interaction system was done using a 2D displayed avatar and speech interface. More work is needed in order to make the SEMAINE API fully suitable for a broad range of applications in the area of emotion-aware systems [3].

Emotion recognition, in robotics context, is the capability of automatically recognizing which emotion a human is expressing among a finite scope of possibilities; this can be done using one or more modalities. In our case the scope is {*neutral*, *happy*, *sad*, *fear*, *anger*} and our modalities are image and sound.

Classifying emotions in human conversation was studied in [4] where it was presented a comparison between various acoustic feature sets and classification algorithms for classifying spoken utterances based on the emotional state of the speaker. Later in [5] was presented an *emotion recognition* system to classify a human emotional state from audiovisual signals. The strategy was to extract prosodic, Mel-Frequency Cepstral Coefficient (MFCC), and formant frequency features to represent the audio characteristics of the emotional speech.

A face feature extraction scheme based on HSV color model was used to detect the face from the background. The facial expressions were represented by Gabor wavelet features. This proposed emotional recognition system was tested and had an overall recognition accuracy of 82.14% of true positives. Recently in [6] it was described a multi-cue, dynamic approach to detect emotions in video sequences. Recognition was achieved via a recurrent neural network, whose short term memory and approximation capabilities are appropriate for modeling dynamical events in facial and prosodic expressivity.

As state-of-the-art shows, classifying emotions is a problem that needs to be addressed from several different modalities used by humans in natural communication. When two or more modalities are used and fused as input to the final decision, the system is called a multimodal system. When a multimodal system is devoted to interaction, this system will perform what is called a *multimodal interaction*. Darwin studied *multimodal interaction* in humans in [7], a study about how humans express their emotional states placing a great emphasis on facial expressions. Paul Ekman, using a more modern approach, studied emotional states and facial expressions across cultures [8–10].

2.2 Emotional States

Across human history, emotions have always been considered important; especially by their role in social behavior. Sometimes they were seen as elevating, and in some other times as being degrading. But until recently it was hard to include them in the field of science, and even harder to include them in the realm of technology. However, we can start to address these questions now. For one thing, we have a workable idea of what emotions are and that is a first step in the attempt to discover *why* emotions are and *what* emotions do for us. For another, we know that emotions play a critical role in social behavior [11, 12].

Spinozza [13], during the seventeenth century, worked extensively on an attempt to define what human emotions are. This work was then continued by William James, and more recently by Damasio [11, 12]. But despite all the advancements in neuroscience's understanding of what emotions are, all the work done pointed away from the possibility of mathematically representing emotions. In order to include emotions in technology, it is necessary to find a mathematical framework for emotions. To do that, it must first be clarified that in neuroscience a sharp distinction is made between emotions and feelings.

While emotions are occurrences in our body (including facial expressions), the feelings are the neuronal representations of such emotions. We can see people's emotions, but not their feelings. In most circumstances emotions can generate feelings, but not the other way around. What feelings often elicit is the occurrence of a simulation of an emotion induced by a feeling, an as-if emotion. However, it is possible to make use of the person's expressed emotion to guess the person's implied feeling. So what the robot will do is not to have an emotion, but rather learn the person's implied feeling and how that feeling should influence the robot's decision-making.

The role of feelings in decision-making is that they highlight certain possibilities as being valid, while discarding

other possibilities as being invalid. This discarding of possibilities can be understood as a constraining of the flow of information that goes from axioms/assumptions to statements/decisions. The information based approach to axiomatic systems in mathematics was developed by Chaitin and was based on Leibniz's approach to causality in the universe [14]. While Newton and Spinoza assumed that the past was enough to determine the future, Leibniz assumed that the universe had a certain irreducibility to it which enabled it to have multiple future-possibilities given a certain past. In situations where one does not have the needed information about the past, the use of Leibniz's perspective is often a better approach than Newton's or Spinoza's. Although Damasio's model of emotions was found to be a good match to Spinoza's perspective [11], it might even be a better match to Leibniz's perspective. We will use here a Leibniz-based approach to Damasio's model of emotions proposed in [15].

Newton defined that one cause leads to one consequence, and every consequence is univocally and completely defined by its cause. Leibniz had a more freedom-occurring approach to causality, believing that the cause limits the possible consequences but not to the point of there being only one consequence. Lori [16] explored the point of view of Leibniz using the approach to Leibniz's causality defended in [14] and proposed that the word "causality" be replaced by the concept of "enablement", and the word "consequence" be replaced by the concept of "alternatives". The enablement does not have the capacity to univocally define a single alternative because the amount of information, the "message", coming into a system is not capable of doing such a stringent constraining of the alternatives. The enablement can be represented as the probability distribution of a cause (stimulus) generating a certain event (posterior). The Leibniz perspective of causality can be summed up, as a structural triplet approach [16], on which the flow of information goes into an enablement that then generates alternatives, with only a portion of those alternatives being capable of transmitting information. The enablement, the alternatives and the message/information are the three components of the Leibniz approach to causality. In the Leibniz-Damasio model of emotions/feelings, the feelings represent the fulfillment (positive feelings) or the failure (negative feelings) of a prediction. Using the Leibniz approach the negative feelings are divided into three types of failure, one for each of the components of the causality: *de-enabled*, *de-alternatived* and *de-messaged.*

A fourth group, associated to the *success* of the prediction, is also required. But as feelings are about prediction outcome, then only the success-associated emotion is a true feeling, with the negative feelings serving merely as indicators of how far one is from the positive feeling. This approach strongly coincides with the approach of positive psychology where the *Eros vs. Tanatos* duality is abandoned, in favor of a learning of how to cope with the increases and decreases in positive feelings that are a natural occurrence in life, e.g. [17]. In Damasio's perspective there were different types of positive feelings, each of them associated to a certain group of negative feelings (typically 3–4). What is proposed in the Leibniz-Damasio model of feelings is that the different positive feelings can be assigned to different self-consciousness perspectives, with each positive feeling being associated to exactly three negative feelings [16].

The three negative feelings being respectively associated to the *de-enabled*, *de-alternatived* or *de-messaged* forms of failing to attain the positive feeling for that level of self-consciousness. The different self-consciousness levels considered in the Leibniz-Damasio approach are first those proposed by Damasio: *proto-self*, *core consciousness*, and *extended consciousness*. But in Damasio's work, extended consciousness refers only to consciousness about one's life as an individual, an autobiographical self-consciousness, so we simply call it Personal-consciousness. The Leibniz-Damasio approach to feelings further considers two other types of extended consciousness: Historical-consciousness, and Universal consciousness. In Historical-consciousness the self is extended to the life of the person's culture (in both its social and religious aspect), and in Universal-consciousness it is further extended to the life of the whole Universe the person inhabits. The Leibniz-Damasio approach accounts for all of the social emotions/feelings in Damasio's approach, and puts them inside a Leibnizian perspective. In this work, we only considered the Core-consciousness level. Each positive feeling is associated to an Emotional Competent Stimulus (ECS) for the corresponding emotion. Damasio did not define ECS for the neutral state, as people are typically feeling something. But one can consider that feelings below a certain threshold of intensity can be considered as non-existent, and in that sense we can propose the addition of a fifth group where the neutral state is.

The recent research by Damasio [18] agrees with his previous research about the different types of consciousness, and gives further strength to the possible relation to Leibniz causality in at least the following ways: (a) It considers that feelings of emotions are a composite of perceptions of perceived body states (the message), with altered mental script deployment (the alternatives); with that composition occurring in different parts of the brain (the enablement). (b) The feelings of emotions are constructed based on primordial feelings, with a mechanism of such construction being interoception; meaning that the self-concept is a major component of feelings in agreement with [16] and Table 1. (c) The level of the autobiographical self-consciousness is used as a definer of the difference between situations where we simply have to take into consideration the consequences to oneself, versus the situations where one needs to take into account the consequences to those around us; the importance and the

Table 1 Relation between Damasio's condition, consciousness level and human emotions

Condition	Consciousness level		
	Proto-self	*Core*	Personal
De-alternative	Tension	*Anger*	Disgust
De-enabled	Fatigue	*Fear*	Surprise
De-message	Malaise	*Sad*	Jealousy
Successful	Well-being	*Happy*	Pride
	Neutral		

moment-to-moment fluency of this level is quite compatible with the fluency of feeling and is thus in agreement with our proposal that different feelings are associated to different consciousness levels, as expressed in [16] and Table 1.

Table 1 shows Leibniz-Damasio's levels of consciousness, with its four feelings per consciousness level [16], and also the neutral state proposed here. In this paper we will only consider the feelings associated to the Core-consciousness level (anger, fear, happiness, sadness, neutral state). The Emotional Competent Stimulus for each one of these four emotions are:

- *Neutral*—the absence of an emotional competent stimulus is what we consider as a trigger for the *neutral* state.
- *Happy*—recognition (in others or in self) of a contribution to cooperation and/or communication. This emotion is associated with successful cooperation and/or communication between self and another individual, and so it is linked to the *successful* emotions.
- *Sad*—individual suffering/in-need. This emotion is associated with reduction/loss of the capacity to communicate with an individual, and so it is linked to the *de-messaged* emotions.
- *Fear*—weakness/failure/violation of the individual's own person or behavior. This emotion is associated with de-empowerment of the individual, and so it is linked to the *de-enabled* emotions.
- *Anger*—an interlocutor's violation of norms. This emotion is associated with loss of alternative possibilities of communion and/or cooperation, and so it is linked to *de-alternative* emotions.

So the first step in this direction is to endow the robot with the capacity to analyze the emotions of the human.

The emotional state that a person demonstrates as a reaction to some circumstance depends on the *social behavior profile* of the subject. The definition of *Social behavior profile* (SBP) is context dependent, but frequently it is a way of defining a scope for personality. In [19], it is discussed that both sympathy and antipathy can but do not need to be empathic, along [19] both antipathetic and sympathetic are considered as *social behavior profiles*. Moreover in [20] an attempt to do automatic analysis of learner's social behavior during computer-mediated synchronous conversations was presented. Four SBPs were analyzed in that work: moderator, valuator, seeker, interdependent. In medicine, for example on [21], we found autism, apathetic and aggressiveness as *social behavior profiles*. Several other *social behavior profiles* are also listed in the literature, but they may vary considerably depending on the author's interpretation and on the context of each problem. For our context, we selected three *social behavior profiles* for our robot: sympathetic, antipathetic and humorous.

2.3 Probabilistic Approach to Deal with Uncertainty

In Bayesian algorithms, it is important to define a technique to fill-out the *Bayesian network* with information from the real world. Learning techniques are widely used for designing and testing natural language processing systems [22]. A particular case of learning techniques was discussed in [23], where the problem of spoken interaction was addressed. In our system, during the learning phase, the human experimenter embodies the strategy of the robot and interacts with another human. Meanwhile, the robot *looks* and *listens* to these two humans interacting. From the observation, the desired variables (that in our case are defined on Sects. 3.1 and 4.1) were extracted and the *Bayesian network* was filled. At the running phase, when the decision moment arrives, the robot had a state and also a filled *Bayesian network*, thus it inferred and performed the correct responses.

2.4 Bayesian Framework Applied to Our Context

Our objective was to create a smart robotic system with contactless interfaces (cameras and microphones) capable of a multimodal interaction with a human. Phoneme recognition was not our concern, we were dealing with *emotion recognition* only on the auditory part, thus a story board for the human input was used, nevertheless the robot responses varied.

Figure 2, adapted from [24], shows a conceptual drawing of the system, which is composed by three main parts: *memory* (where the knowledge acquired during the learning phase and also over time was stored), *analysis and synthesis.*

Notice that there is a pair of *analysis* parts (from stimulus to Posterior), one per each channel. Later the visual and auditory channels information merged into the decision rule, where both posteriors were taken into account. An overview of the *analysis* part is described bellow:

- Stimulus: in our case, both image and sound.
- Sensory Processing: it is here that the facial Action Units (*AUs*) and the auditory variables are identified from the raw stimulus. This leads to the sensory input for the Bayes Rule ((1) explained on Fig. 2 on bottom right corner) to

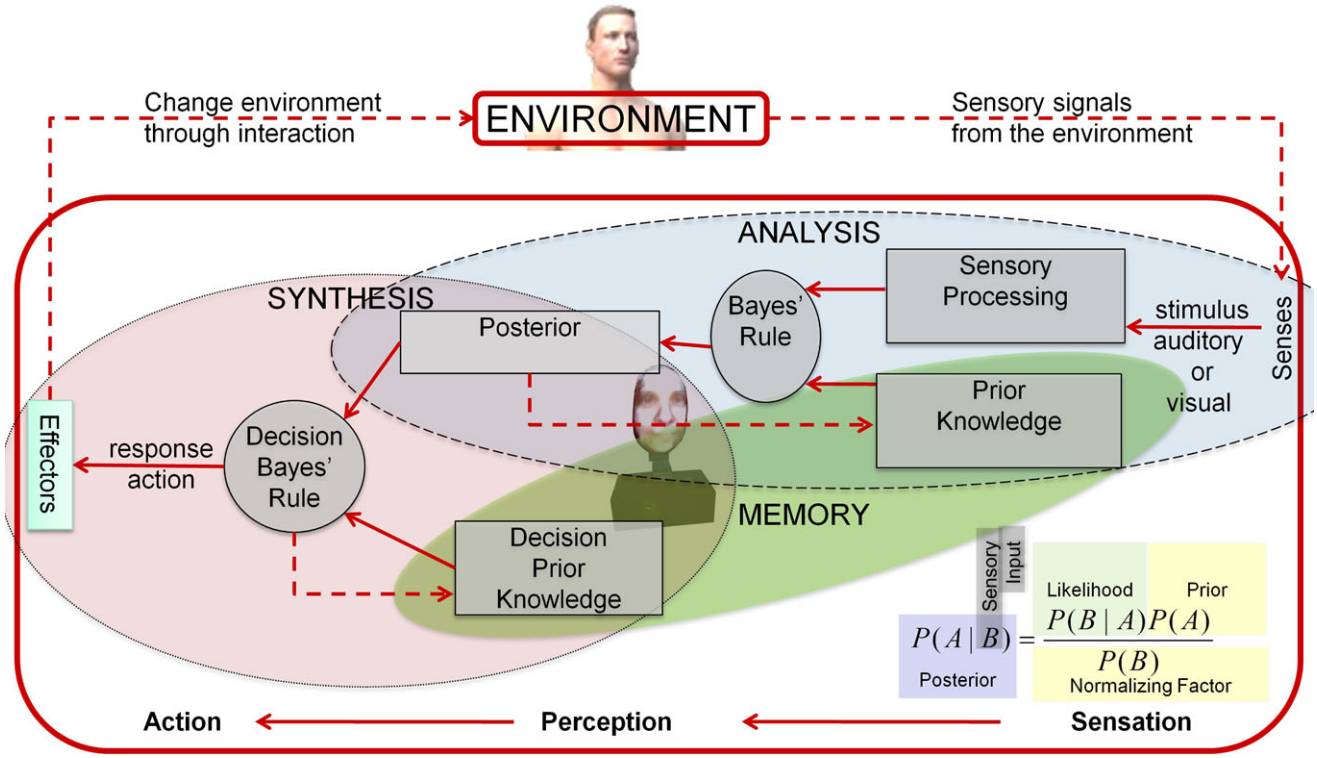

Fig. 2 The analysis ellipse (*dashed*) is composed by two layers: one for visual analysis and the other for auditory analysis. The synthesis ellipse (*round dotted*), where the fusion happens, is a single layer. Per each modality, the stimulus comes from the sensors (camera and microphone), passes through the sensory processing where one group of features is extracted; this group of features is the input for the Bayes's rule which will, based on the prior knowledge, infer the posterior classification. Later the two posteriors and the robot *social behavior profile* (which is hardwired in memory) are merged in the synthesis probabilistic function, and the response (robot emotional state) is passed to the effectors

ask the correct question: "what is the probability of a posterior, given the input from sensory processing?".

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} \quad (1)$$

- Prior Knowledge: contains the prior and the likelihood. According to Bayesian Theory, prior is the probability of each event happening over a stimulus, independent of any other event. Once more according to Bayesian Theory, the likelihood is the probability of an event happening given another event, it is filled out during the learning phase and it is stored in *memory*.
- Bayes's Rule: inference executed over the prior and the likelihood, in order to give the probability of posterior given the input from sensory processing.
- Posterior: inferred result.

In Fig. 2, the arrow that comes from posterior to the prior knowledge indicates that the *Bayesian network* has feedback. In other words, it starts by knowing just the prior, but through the passage of time, the robot acquired more "life experience" and the posterior became part of the prior, thus the probability of an event that was already observed increased.

An overview of the *synthesis* part is described bellow:

- Posterior: the inferred results from the analysis are the inputs for the synthesis.
- Decision Prior Knowledge: this prior knowledge balances the fusion of the modalities.
- Decision Rule: Bayes's rule that infers a final decision over the emotional states coming from the posteriors of both modalities.
- Response: Decision of what the effectors will actually do.

Figure 3 shows the system implementation modules. *Visual perception* is divided between sensory processing (explained on Sect. 3.1) and the classifier that will be detailed on Sect. 3.2. *Auditory perception* sensory processing is explained on Sect. 4.1 and its classifier will be described on Sect. 4.2. The Decision Process is where we do a fusion of the two classifications with a given *social behavior profile* (SBP) for the robot, this will be detailed in Sect. 5.1. After the fusion, the effectors take place doing the synthesis of facial expressions (Sect. 5.2.1) and the synthesis of vocalization (Sect. 5.3).

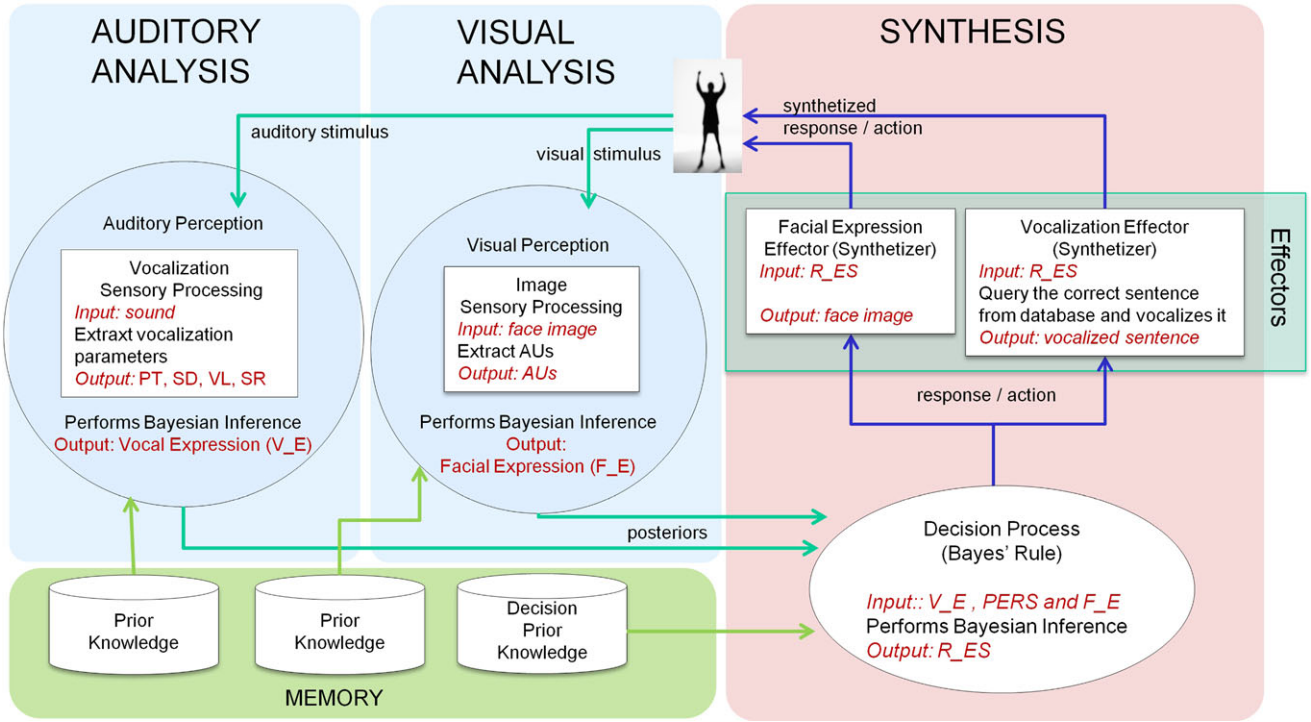

Fig. 3 Lower level schema where the two analysis are clearly separated with their respective input and output, the posteriors of the analysis are V_E and F_E. The decision process (the fusion) also takes into account the given *social behavior profile*. The output of the fusion is *RES* that stands for emotional response. *RES* is then the input for both effectors. The synthesized response is the sound and the synthetic face produced by the robot

3 Image Emotion Recognition

3.1 Image Sensory Processing

The first step in image emotion recognition, before classification was done, is the extraction of features which in our case are the Action Units. In human beings the facial expressions are distortions or movements of facial features (e.g., eyebrows, eyes, nose, mouth) and arise as a result of muscular activity. This activity can be performed voluntarily; for example when performing a grimace. However, human beings systematically perform involuntary facial expressions. These are a form of nonverbal communication used in social contact as means to express emotions. In fact, humans are beings with a strong social characteristic, and facial expressions are a primary mean of conveying social information. In Fig. 4, examples of facial expressions are presented, typically associated to some emotional states.

In human beings, the association between emotional states and facial expressions are so extensive that, for certain emotions. It can be very difficult to avoid doing the characteristic facial expressions; even when one wants to hide the real emotional state. Nevertheless, this close relationship between emotional states and facial expressions may work "in the opposite direction"; it is possible to induce the corresponding emotional state in an interlocutor only by performing, voluntarily, a facial expression.

Fig. 4 Facial expressions associated to interlocutor's emotional states: {*neutral*, *happy*, *sad*, *fear*, *anger*}

Darwin studied how humans express their emotional states [7]. It is an extensive study focusing on the various forms used by humans to express themselves: facial expressions, gestures, vocalization, etc. More recently, Paul Ekman devoted specific attention to the subject of emotional states and facial expressions [8–10]. In [10] it is mentioned that an alternative approach to measuring facial expressions of emotion is through systematically examining video records to identify the muscular movements that constitute the emotional expressions. One of the advantages of this approach is that it is totally unobtrusive.

Facial Action Coding System (FACS) [8] defines a total of 52 Action Units (*AUs*) where 8 of them are related with the head pose. The remainder 44 concern small distortions

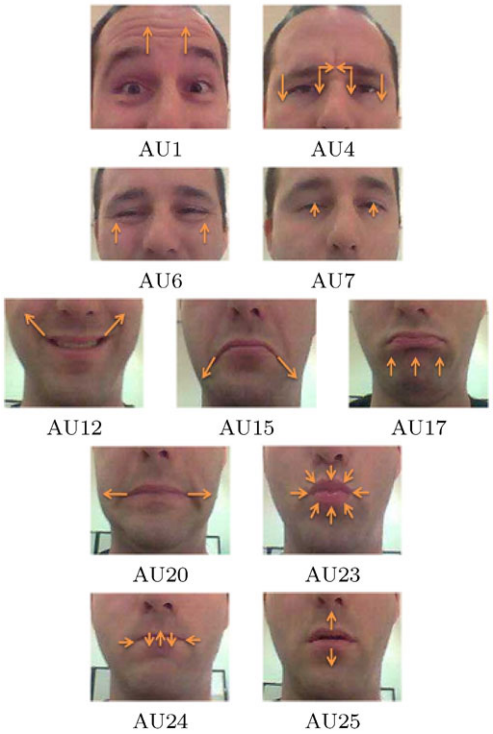

Fig. 5 Examples of Action Units, see description on Table 2

Table 2 Description of Action Units

Action Unit	Description
*AU*1	Inner portion of the brows raised
*AU*4	Brows lowered and drawn together
*AU*6	Cheeks raised
*AU*7	Lower eyelids raised
*AU*12	Lip corners pulled obliquely
*AU*15	Lip corners pulled down
*AU*17	Chin boss pushed upwards
*AU*20	Mouth stretched horizontally
*AU*23	Lips tightened
*AU*24	Lips pressed together
*AU*25	Lips relaxed and parted

over the face which characterize the facial expressions. Each of these *AUs* is anatomically related to the activity of a specific set of muscles which produces changes in the facial appearance. Therefore, a facial expression can be interpreted as a set of specific *AUs*, which causes "distortions" in facial features (i.e., mouth, eyes, eyebrows or noise). By identifying these distortions, facial expressions can be recognized. In our work, only a small sub-set of the *AUs* introduced by Ekman was used (examples are presented in Fig. 5).

We can summarize our visual feature extraction by the following steps:

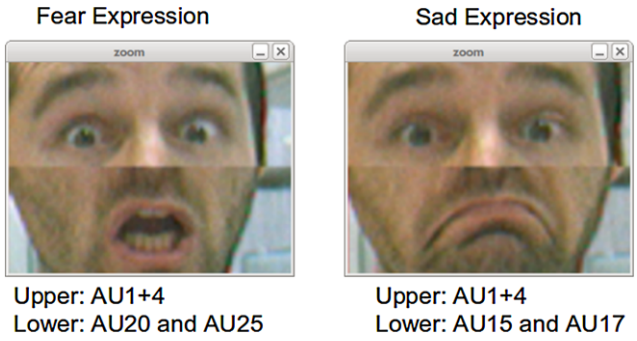

Fig. 6 Action Unit feature extraction example results. On the *left* image is a fear expression being performed; on the *right* is a sad expression. On both images, notice that the face is divided into upper and lower face while the rest is ignored. Below each of the images we can see the detected *AUs*

Step 1: The human performs a facial expression in front of the camera.
Step 2: Each frame of the video is recorded.
Step 3: We need to detect a human face from each frame. We are using OpenCV well-known haarlike features to do that. After the face is detected, it is divided in upper face and lower face.
Step 4: Then we detect some, among the 13 possible evidences from each face: *AU*1 *AU*4 AU1 + 4 *AU*6 *AU*7 *AU*12 *AU*15 *AU*17 *AU*20 *AU*23 *AU*24 *AU*25.

Thus, we implemented our own method to detect this *AUs*. The method implemented is based on PCA (Principal Component Analysis). The upper face and lower face are treated independently, a feature vector is then extracted from each image while the user performs that specific Action Unit. After this, we have information enough to detect whether the Action Unit is present or not in the upper face. The same was done for the lower face, independently of the result for the upper face. This test was done for all the 13 possible *AUs*. The result of this method is binary, the *AU* was either present or not present. Some results can be seen in Fig. 6.

3.2 Classification of Facial Expressions

Once the feature extraction problem is solved, our robot must then classify the facial expressions. In this case only five possibilities were considered: {*neutral*, *happy*, *sad*, *fear*, *anger*}. A *Bayesian network* (Fig. 7) was used to classify the facial expressions performed by the human interlocutor. FACS [8] was used as a theoretical basis to design the classifier. The Action Units we considered as associated to each one of these facial expressions are shown in Table 3.

When designing the classifier, we made our own interpretation of FACS, which drives a set of random variables different from those defined by other researchers. Each facial expression is composed by a specific set of Action Units.

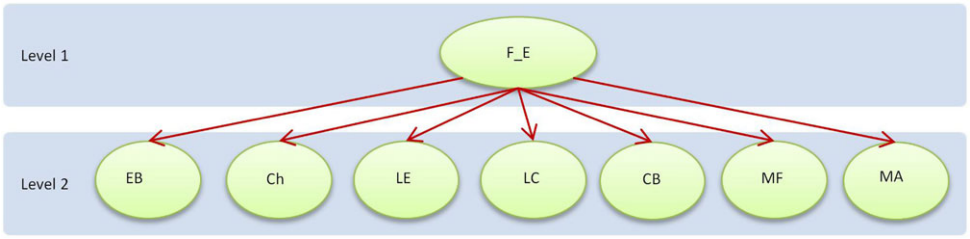

Fig. 7 Facial expression dynamic *Bayesian network*. This is the DBN1 mentioned on the Introduction and also represented in Fig. 1

Table 3 Discrimination of the *AUs* that are present in every one of the facial expressions

	Upper face			Lower face			
	Eyebrows	Cheeks	Lower eyelids	Lip corners	Chin boss	Mouth form	Mouth aperture
Neutral	–	–	–	–	–	–	–
Happy	–	*AU*6	–	*AU*12	–	–	*AU*25
Sad	*AU*1 + 4	–	–	*AU*15	*AU*17	–	–
Fear	*AU*1 + 4	–	–	–	–	*AU*20	*AU*25
Anger	*AU*4	–	*AU*7	–	*AU*17	*AU*23	*AU*24

Every one of these Action Units is a distortion of a facial feature induced by muscle activity. Normally, a well determined set of muscles is associated to a specific Action Unit, which can give the idea that all these basic distortions are independent. Nevertheless, some of these Action Units are antagonistic. One concrete, and understandable, example is the case of two Action Units related with the movements of the corners of the mouth, that is *AU*12 and *AU*15. When performing the first one of these, the lip corners were pulled obliquely in the direction of the ears and eyes; meaning that the corners move up and back. By the contrary, when performing just the *AU*15 the lip corners were pulled down. Therefore, if by one way the movements of the lip corners can be considered independent because they are performed by distinct muscle sets, by another, when analyzed visually they are antagonistic, exclusive and non-independent.

Nevertheless, sometimes Action Units seemingly antagonistic and mutually exclusive can occur simultaneously; in this case the term used to describe this situation is "non-additive combination". One example of this situation occurs sometimes when a human being is shown a facial expression of sadness. In these cases, the set of muscles responsible for the *AU*1 were activated together with another set responsible for the *AU*4. In terms of appearance, when *AU*1 occurs alone the inner eyebrows were pulled upwards and, when *AU*4 occurs alone the eyebrows were pulled together and downwards. Therefore *AU*1 and *AU*4 are antagonistic. In reality, it was possible the activation of the two sets of muscles and in this situation we were in the presence of a "non-additive combination". In that case the notation used is *AU*1 + 4 (it is different from the notation *AU*1 + *AU*4, which would be used if these *AUs* could appear in an "additive combination").

Based in these principles, belief variables were defined and a Bayesian classifier of facial expressions was developed; which is described in the next subsection.

3.2.1 *Facial Expression Bayesian Network*

To classify the facial expressions performed by the human being dialoguing with a robot, a *Bayesian network* was developed. The structure of this network of two levels is illustrated in Fig. 7.

In the *Bayesian network*'s first level there is only one node. The global classification result obtained is provided by the belief variable associated with this node: $F\_E \in \{anger, fear, happy, sad, neutral\}$, where the variable name stands from Facial Expression. Considering the structure of the *Bayesian network*, the variables in their second level have as parent this one in the first level: F_E.

In the second level there are seven belief variables:

- $EB \in \{AU1, AU4, AU1 + 4, none\}$ is a belief variable related with the *Eye-Brows* movements. The three events are directly related to the existence of *AU*1, and *AU*4 alone or together (in this case a distinct event was created because it is a "non-additive combination").
- $Ch \in \{AU6, none\}$ is a belief variable which is related with *Cheeks* movements; more specifically, the events indicate if the cheeks are raised (*AU*6 is performed).
- $LE \in \{AU7, none\}$ is a belief variable which is related with the *Lower Eyelids* movements; *AU*7 is the action unit associated with the raising of the lower eyelids.
- $LC \in \{AU12, AU15, none\}$ is the belief variable associated with the movements of the *Lips Corners*. The event *none* must have a high probability when the corners did not perform any movement. The event *AU*12 must have a

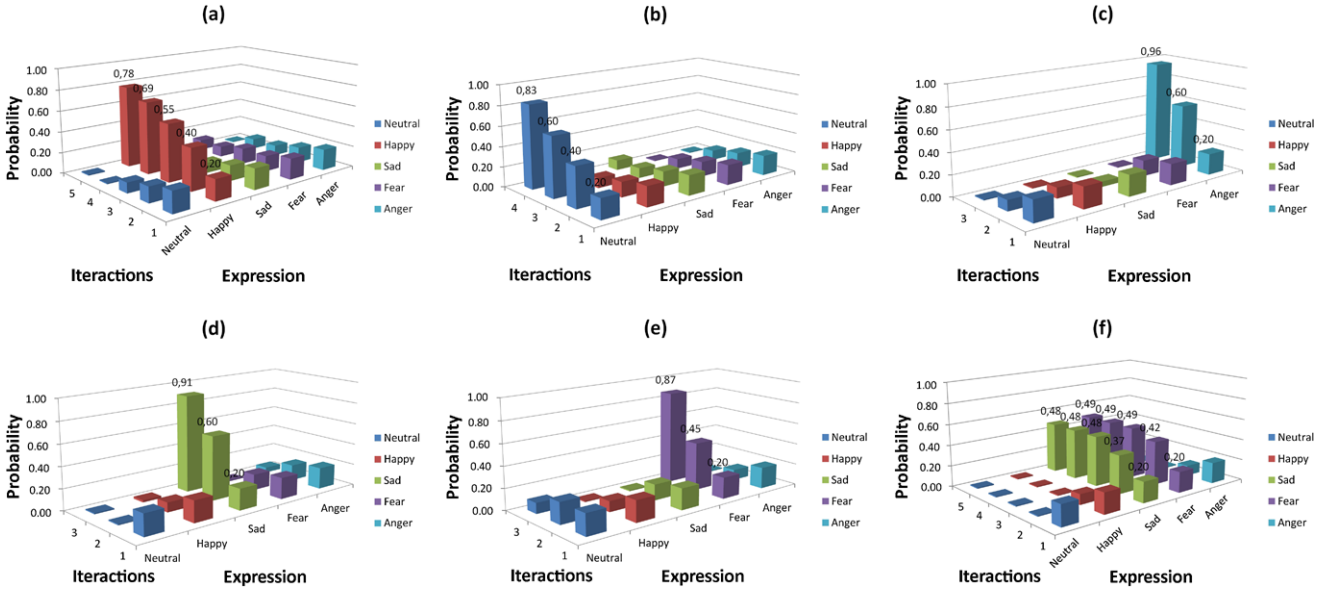

Fig. 8 Results from facial expression classifier: camera grabbing was set to 5 fps, therefore, the iteration axis represents the 5 (or less) utterances that happens inside one second. The expression axis is the selected scope of possible expressions. Notice that the sum of probability at each iteration among the five possible expressions is always 1. In examples (**a**), (**b**), (**c**), (**d**) and (**e**), respectively, inputs were given for *happy*, *neutral*, *anger*, *sad* and *fear*; the dynamic Bayesian network was capable of classifying the expected expression with a fast convergence. In (**f**), an example of ambiguity and misclassification is shown, where the expected result was *sad* but the result of classification was *fear*

great probability when the lip corners are pulled obliquely up and backwards. If the lip corners moved downwards the event $AU15$ must have a great probability.

- $CB \in \{AU17, none\}$ is the belief variable collecting the probabilities related with the *Chin Boss* movements. The event *none* is related with the absence of any movement, while the event $AU17$ had a great probability when the chin boss is pushed upwards.
- $MF \in \{AU20, AU23, none\}$ is the belief variable associated with the Mouth's Form. The events $AU20$ and $AU23$ indicated, respectively, if the mouth is stretched horizontally or, inversely, if the lips are tightened.
- $MA \in \{AU24, AU25, none\}$ is the belief variable associated with the Mouth's Aperture. The events $AU24$ and $AU25$ are related, respectively, with lips pressed together or with lips relaxed and parted.

The following equations illustrate the joint distribution associated to the Bayesian Facial Expressions Classifier.

$$\begin{aligned} &P(F\_E, EB, Ch, LE, LC, CB, MF, MA) \\ &\quad = P(EB, Ch, LE, LC, CB, MF, MA|F\_E) * P(F\_E) \\ &\quad = P(EB|F\_E) * P(Ch|F\_E) * P(LE|F\_E) \\ &\qquad * P(LC|F\_E) * P(CB|F\_E) * P(MF|F\_E) \\ &\qquad * P(MA|F\_E) * P(F\_E) \end{aligned} \tag{2}$$

The last equality is written assuming that the belief variables in the second level of the *Bayesian network* are independent.

From the joint distribution, the *posterior* can be obtained by the application of the Bayes rule as follows:

$$\begin{aligned} &P(F\_E|EB, Ch, LE, LC, CB, MF, MA) \\ &\quad = \frac{P(F\_E, EB, Ch, LE, LC, CB, MF, MA)}{P(EB, Ch, LE, LC, CB, MF, MA)} \\ &\quad \propto P(EB|F\_E) * P(Ch|F\_E) * P(LE|F\_E) \\ &\qquad * P(LC|F\_E) * P(CB|F\_E) * P(MF|F\_E) \\ &\qquad * P(MA|F\_E) * P(F\_E) \end{aligned} \tag{3}$$

3.2.2 Results of Facial Expressions Bayesian Network

Expected results for the *analysis* part are a correct classification of facial and vocal expressions according to what is expected. Convergence is also expected to appear as time passes, since both *Bayesian Networks* are Dynamic. Figure 8 shows results of Bayesian inference for the facial expressions classifier.

Figure 8(a) shows an experiment that took 5 iterations with the following constant evidences: $EB = none$, $Ch = AU6$, $LE = none$, $LC = AU12$, $CB = none$, $MF = none$, $MA = AU25$. Notice that the convergence happened fast, after the second iteration the best result was already visible.

The classification was considered completed when the percentage was higher than 80% for one of the expressions, or when it reached 5 iterations. Usually the convergence happened in less than 5 iterations, like in examples of Figs. 8(b), (c), (d) and (e) where the inputs were respectively:

- (b) "$EB = none$, $Ch = none$, $LE = none$, $LC = none$, $CB = none$, $MF = none$, $MA = none$";
- (c) "$EB = AU4$, $Ch = none$, $LE = AU7$, $LC = none$, $CB = AU17$, $MF = AU23$, $MA = AU24$";
- (d) "$EB = AU1+4$, $Ch = none$, $LE = none$, $LC = AU15$, $CB = AU17$, $MF = none$, $MA = none$";
- (e) "$EB = AU1+4$, $Ch = none$, $LE = none$, $LC = none$, $CB = none$, $MF = AU20$, $MA = AU25$".

A misclassification is presented on Fig. 8(f), the expected expression was "*sad*", however it was a case where the *sensory processing* phase failed, thus it became ambiguous between "*sad*" and "*fear*" and the result was a misclassification to "*fear*".

4 Audio Emotion Recognition

4.1 Audio Sensory Processing

The first step in audio emotion recognition, before classification, was feature extraction. When a human speaks a phrase on the microphone, each phrase will be recorded on a wav file (we used wave files of the format wav: the most common wav format contains uncompressed audio in the linear pulse code modulation format). From this wav file it is possible to detect some parameters that arise as a result of the wave characteristics. These characteristics are involuntarily performed by the human when a certain emotion affects the voice.

We can summarize our audio feature extraction on the following steps:

Step 1: The human spoke a phrase on the microphone.
Step 2: Each second of the phrase was recorded on a different wav file.
Step 3: The phrase was then mounted together and re-recorded in a wav file.
Step 4: It was necessary to detect 3 variables from each wav file.

These variables are described on Table 4.

We used the Praat toolkit [26] to detect these evidences (see example in Fig. 9). A vocal expression was then the result of processing these features using our Bayesian classifier that will be explained on Sect. 4.2. A vocal expression can be in the scope of {*neutral*, *happy*, *sad*, *fear*, *anger*}.

Table 4 Description of variables extracted from sound

Variable	Description
SD	Since we know the sampling frequency (*sfreq*) of the acquired sound, we also know the beginning and the end of each sentence, and consequently the number of samples (*nsam*); and then it is simple to determine the duration in seconds by $SD = nsam/sfreq$.
PT	Stands for *pitch*, pitch represents the perceived fundamental frequency of a sound. The pitch extraction was done by autocorrelation method [25].
VL	Stands for Volume Level. This variable is actually the *energy* or *intensity* of the signal, which for a theorically continuous-time signal $x(t)$ is given by $VL = \int \lvert x(t)^2 \rvert dt$.

4.2 Classification of Vocal Expressions

In *auditory perception*, after the feature extraction (see Sect. 4.1) done in sensory processing (see Fig. 3), the classifier takes these audio features as input and then classified a vocal expression. Here we are going to explain how this information was processed into our *auditory perception Bayesian network*. The robot needed to be capable of classifying among the possible Vocal Expressions: {*neutral*, *happy*, *sad*, *fear*, *anger*}.

4.2.1 Auditory Perception Bayesian Network

To classify the vocal expressions done by the human while interacting with the robot, a *Bayesian network* was developed by us. The structure of this network of two levels is illustrated in Fig. 10.

In the *Bayesian network*'s first level there was only one node. Furthermore, the global classification result obtained was provided by the belief variable associated with this node: $V\_E \in \{neutral, happy, sad, fear, anger\}$, where the variable name stands for Vocal Expression. Considering the structure of the *Bayesian network*, the variables in their second level have as parent this one in the first level: V_E.

Although the second level has only three belief variables, the scope of each of these variables is big.

- $PT \in \{from\ 75\ to\ 6000\ Hz\}$ was a belief variable related with the *Pitch*. The three events were directly related with the perceived frequency of the signal duration of the interlocutor's phrase.
- $SD \in \{from\ 0\ to\ 10\ seconds\}$ was a belief variable which was related with *phrase Duration*.

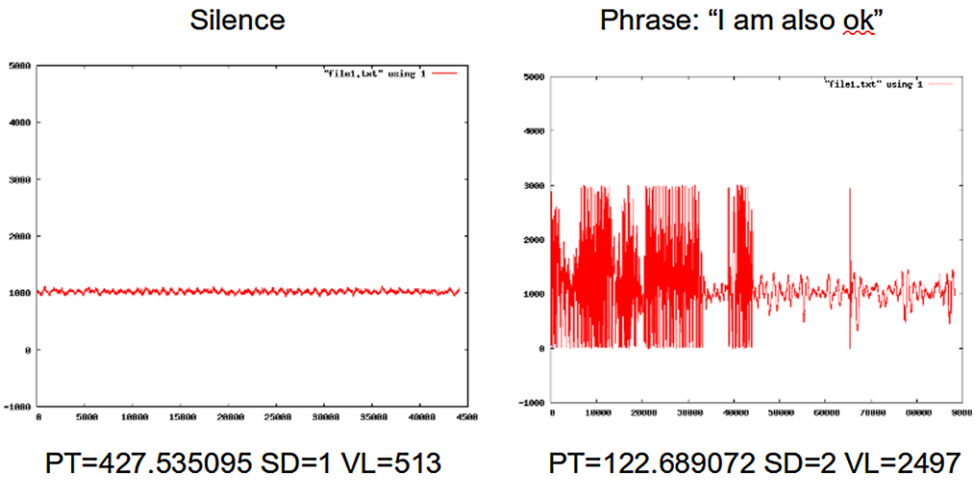

Fig. 9 Sound feature extraction example results. On the left image is one second of silence; on the right is a two seconds phrase where it is said "I am also ok". On both images, notice that the amplitude of the signal is limited by both a maximum and a minimum thresholds. Below each of the images we can see the detected audio features given by equations and methods referred in Table 4

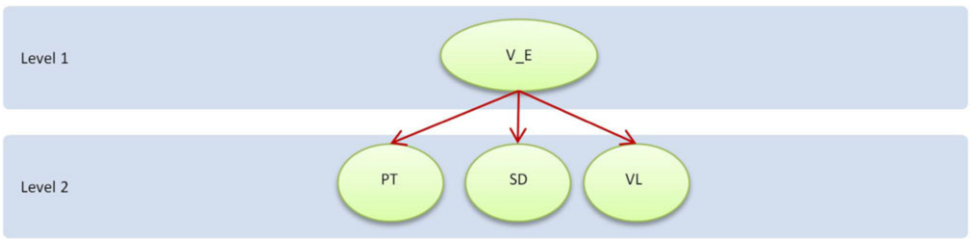

Fig. 10 Dynamic *Bayesian network* for *Auditory Perception*. This is the DBN2 mentioned on the introduction and also represented in Fig. 1

– $VL \in \{from\ 0\ to\ 5000\ arbitrary\ unit\}$ was a belief variable which stood for *Volume Level*, or in other words, it is the energy of the signal.

The following equations illustrate the joint distribution associated to the Bayesian Vocal Expressions Classifier:

$$\begin{aligned} &P(V\_E, PT, SD, VL) \\ &\quad = P(PT, SD, VL|V\_E) * P(V\_E) \\ &\quad = P(PT|V\_E) * P(SD|V\_E) * P(VL|V\_E) * P(V\_E) \end{aligned} \tag{4}$$

The last equality can only be done if it is assumed that belief variables *PT*, *SD* and *VL* are independent.

From the joint distribution, the *posterior* can be obtained by the application of the Bayes Formula as follows:

$$P(V\_E|PT, SD, VL) = \frac{P(PT|V\_E) * P(SD|V\_E) * P(VL|V\_E) * P(V\_E)}{P(PT,\ SD,\ VL)} \tag{5}$$

From the summation theorem we can calculate:

$$\begin{aligned} &P(PT,\ SD,\ VL) \\ &\quad = P(PT|V\_E) * P(SD|V\_E) * P(VL|V\_E) * P(V\_E) \\ &\qquad + P(PT| \sim V\_E) * P(SD| \sim V\_E) \\ &\qquad * P(VL| \sim V\_E) * P(\sim V\_E) \end{aligned} \tag{6}$$

4.2.2 *Results of Bayesian Network for* Auditory Perception

The robot is able to infer over the likelihoods when interacting with the user. The expected results for the *analysis* part are a correct classification of facial and vocal expressions. Convergence is also expected to appear across time, since both *Bayesian Networks* are Dynamic.

Figure 11(b), shows results of the Bayesian inference during 4 iterations with the following constant evidences: $Pitch = 136.569794$, $SentenceDuration = 3$, $VolumeLevel = 1170$. Notice that the convergence happened fast, after the second iteration the best was already visible. Usually the convergence happened in less than 5 iterations, like in examples of Figs. 8(a), (c), (d) and (e) where the inputs were respectively:

– (a) "$Pitch = 140.264309$, $SentenceDuration = 3$, $VolumeLevel = 1260$";
– (c) "$Pitch = 120.473537$, $SentenceDuration = 2$, $VolumeLevel = 2147$";
– (d) "$Pitch = 138.326496$, $SentenceDuration = 2$, $VolumeLevel = 865$";
– (e) "$Pitch = 137.345883$, $SentenceDuration = 4$, $VolumeLevel = 1477$".

A misclassification is presented on Fig. 8(f), the expected expression was "*happy*", however it was a case where the

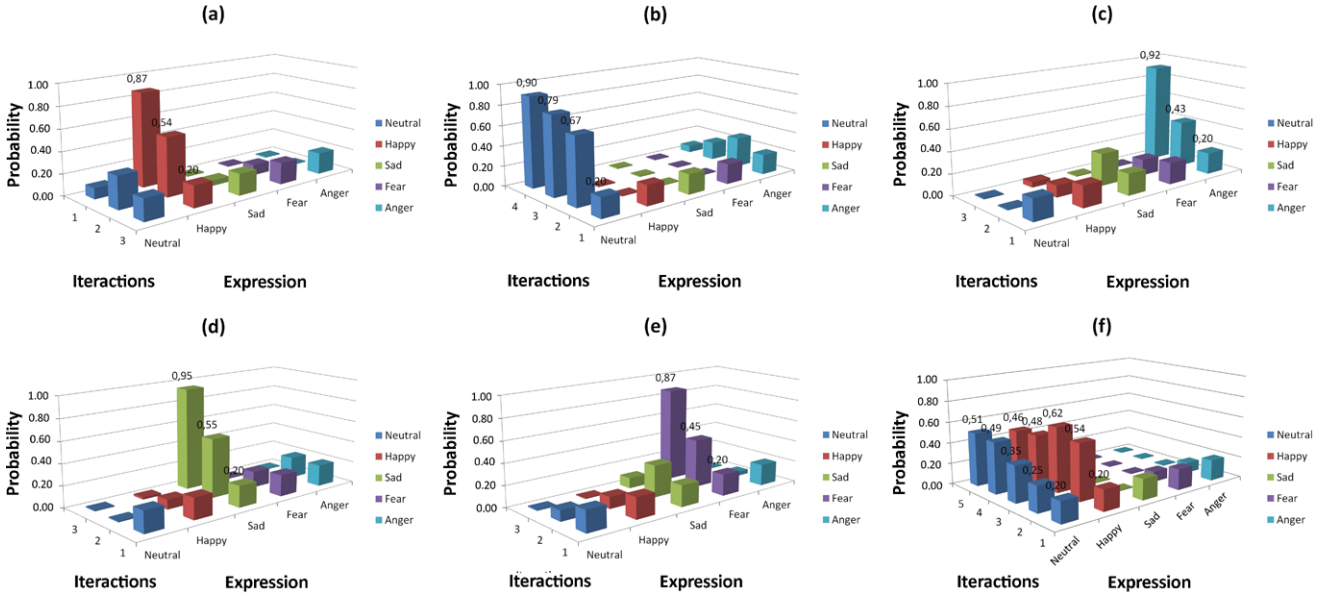

Fig. 11 Results from analysis of vocalization: the sentence sound is recorded and divided second by second, thus, the iteration axis represents the 5 (or less) utterances that happen inside one sentence. The expression axis is the selected scope of possible vocal expressions. The sum of probability at each iteration among the five possible vocal expressions is always 1. In examples (**a**), (**b**), (**c**), (**d**) and (**e**), respectively, inputs were given for *happy*, *neutral*, *anger*, *sad* and *fear*; the dynamic Bayesian network was capable of classifying the expected expression with a fast convergence. In (**f**), an example of ambiguity and misclassification is shown, where the expected result was *happy* but the result of classification was *neutral*

sensory processing phase failed, thus it became ambiguous between "*happy*" and "*neutral*" and the result was a misclassification to "*neutral*".

5 Modeling for Synthesis and Response

For clarity, it must be stated that we consider the fusion to be as much a part of the synthesis as are the effectors. The result of both modalities were combined in this phase and can be used separately or together. The result was a decision of what to synthesize among the possible expressions, which were in the scope {*neutral*, *happy*, *sad*, *fear*, *anger*}. There were clearly nine possible combinations for the system:

1. Analyze *audio* then synthesize *audio;*
2. Analyze *audio* then synthesize *face;*
3. Analyze *face* then synthesize *face;*
4. Analyze *face* then synthesize *audio;*
5. Analyze *audio and face* then synthesize *audio;*
6. Analyze *audio and face* then synthesize *face;*
7. Analyze *audio* then synthesize *audio and face;*
8. Analyze *face* then synthesize *audio and face;*
9. Analyze *audio and face* then synthesize *audio and face.*

Taking advantage of this independence across the modalities, we used option 1 and 3, respectively on the tests presented on Figs. 11 and 8. Henceforth, we focus on option 9 which is the complete fusion.

5.1 Fusion with Social Behavior Profile

According to Fig. 3 the Decision Process received as input F_E (the classified Facial Expression) and V_E (the inferred reaction from *auditory perception* named as Vocal Expression). It took a decision according to these inputs, and according to both the *memory* contents and the given *social behavior profile* ($SBP = \{Emphatic, Antipathetic$ and $Humorous\}$).

For now, we defined 3 possible profiles for the robot: *Sympathetic*, *Antipathetic* and *Humorous.* When the Bayesian inference was executed over the network presented in Fig. 12, the output occurred according to what was trained for each *social behavior profile*.

Figure 12 shows the *Bayesian Network* that does the fusion. It combined V_E (vocal expression), F_E (facial expression) and *SBP* (robot given *social behavior profile*) in order to determine *RES* (the final Response that the robot performs). This fusion implies that the response was based on the core consciousness emotional state perceived by the robot, in a similar way to what Damasio [11] established that humans do.

The classification result obtained by this *Bayesian network* is provided by the belief variable associated with the top node: *RES* {*neutral*, happy*, sad*, fear*, anger**}, where the variable name stands for robotic response; the * represents that this sub-scope is a vector containing a database of possible sentences from all possible *SBPs* for

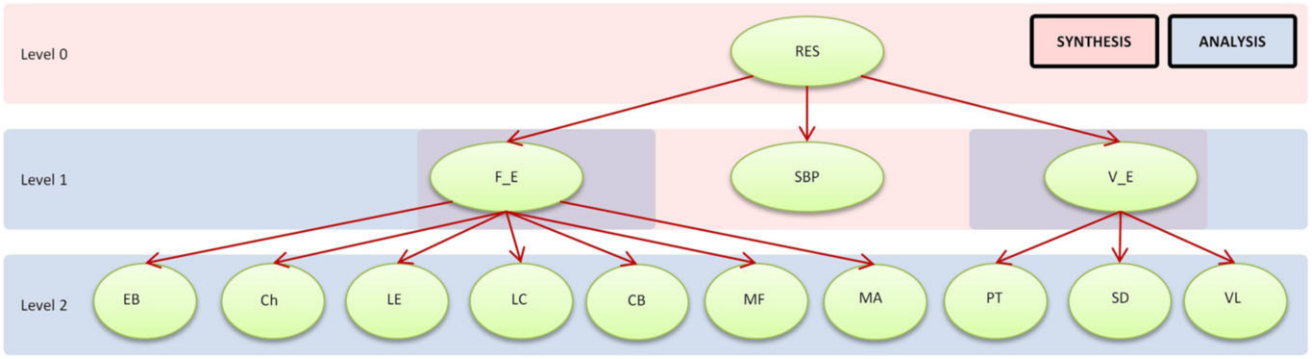

Fig. 12 Dynamic *Bayesian Network* for fusion with robot *social behavior profile*. This is the DBN3 mentioned on the introduction and also represented in Fig. 1

that emotion. Considering the structure of the *Bayesian network*, the variables in their second level have as parent this one in the first level: *RES*.

In the second level there were three belief variables:

- *F_E* {*neutral*, *happy*, *sad*, *fear*, *anger*} represents the facial expression that is given by the previously explained Bayesian classifier.
- *V_E* {*neutral*, *happy*, *sad*, *fear*, *anger*} is a belief variable which represents the vocal expression that is given by the previously explained Bayesian classifier.
- *SBP* {*sympathetic*, *antipathetic* and *humorist*} is a belief variable which stands for *social behavior profile*.

The following equations illustrate the joint distribution associated to the Bayesian Fusion implied by the *social behavior profile*:

$$\begin{aligned} &P(RES, V\_E, F\_E, SBP) \\ &\quad = P(V\_E, F\_E, SBP|RES) * P(RES) \\ &\quad = P(V\_E|RES) * P(F\_E|RES) * P(SBP|RES) \\ &\qquad * P(RES) \end{aligned} \tag{7}$$

The last equality can only be done if it is assumed that belief variables *V_E*, *F_E* and *SBP* are independent.

From the joint distribution, the *posterior* can be obtained by the application of the Bayes Formula as follows:

$$\begin{aligned} &P(RES|V\_E, F\_E, SBP) \\ &\quad = \frac{P(V\_E|RES) * P(F\_E|RES) * P(SBP|RES) * P(RES)}{P(V\_E,\ F\_E,\ SBP)} \end{aligned} \tag{8}$$

5.2 Effectors

5.2.1 Facial Expression Effector

The purpose of the facial expression effector is to produce as output an artificial face image. In our approach, this face

Fig. 13 Facial expression synthesis: {*neutral*, *happy*, *sad*, *fear*, *anger*}

Fig. 14 Facial expression morphing: from neutral (anger = 0%) to angry (anger = 100%)

was human-like and it should be able to produce the five emotional states we covered. The facial expression effector used was also capable of using different input head models that were previously generated from several subjects. Each head model was characterized for its particular face and head shape. Examples of different faces performing the five covered expressions are presented in Fig. 13.

Models of human faces were created, and they were morphed according to *RES* previously defined by the fusion. These models were mesh files which were generated based on three pictures from a person. The Face-Gen 3D head modeler [27] was used to create the meshes. Later these meshes were imported to our OpenCV [28] application and we synthesized expressions. The expression of a face is probabilistic; e.g., the anger level may vary from 0% to 100% as can be seen in Fig. 14.

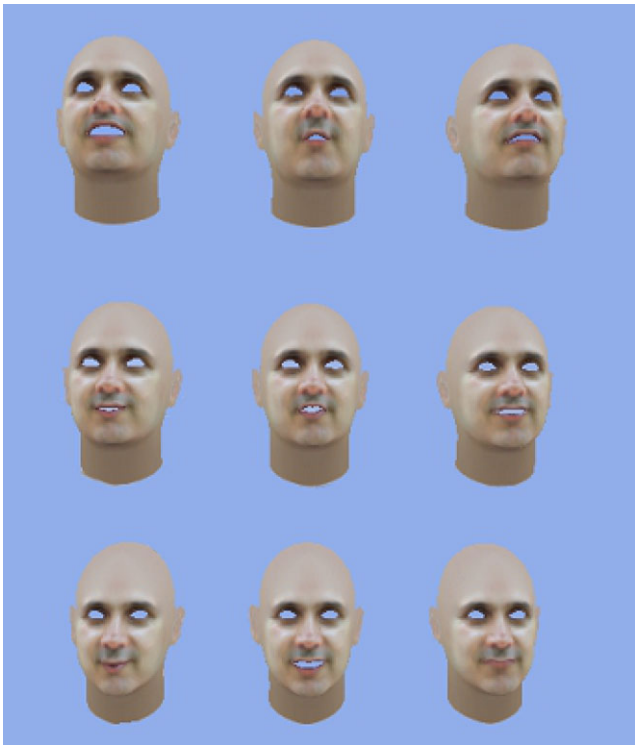

Fig. 15 Nine visemes, each viseme is associated to a phoneme. In this figure, from top-left corner to bottom-right corner, the associated phonemes are: eee, oh, fv, er, YchJ, i, Wu, Ay, and MBP

5.3 Vocalization Effector

The input for Synthesis of Vocalization was the same as for the Synthesis of Facial Expressions (*RES*). However, here the vocalization synthesizer takes this input and continues the story board producing the desired output sound. The vocal expressions phrase database was a previously prepared database of a finite set of possible phrases that can be spoken.

We have also implemented lips synchronization on the avatar, nine visemes that can be seen in Fig. 15, associated to nine phonemes which were used according to what the avatar would speak. Since we were not doing phoneme recognition, this lips synchronization was only possible on the avatar responses where the phrases were known and not on the avatar which was mimetizing the human.

This visemes were associated with phonemes for the English language according to:

1. M, B, P.
2. EEE (long "E" sound as in "Sheep").
3. Err (As in "Earth", "Fur", "Long-er"—also covers phonemes like "H" in "Hello").
4. Eye, Ay (As in "Fly" and "Stay").
5. i (Short "I" as in "it", "Fit").
6. Oh (As in "Slow").
7. OOO, W (As in "Moo" and "Went").
8. Y, Ch, J ("You, Chew, Jalopy").
9. F, V.

6 Autonomy and Intelligence

6.1 Modeling for Memory

Memory was composed of all knowledge that is stored over the lifetime of the system. One portion of memory dedicated to priors was manually filled before the system starts. A prior was what was believed to be the initial probabilities for a *Bayesian network*. Usually it was assumed to be an uniform distribution. Another portion of memory was dedicated to learning Bayesian classifiers of both visual and auditory channels, namely, the likelihood was completed through learning.

Learning was done by putting the system to run in a non-autonomous fashion while gathering the variables' values. During the learning phase, a human expert takes the "correct" decision for the robot while the variables' values were gathered.

We had two macro phases for learning: learning for analysis and learning for decision or synthesis. The learning for analysis was separated for the facial expressions' classification and for the vocal expressions' reaction. The learning for synthesis was just one, since the same learning rule applied for both modalities together and one served as feedback to the other.

Concerning the *learning for Bayesian analysis of facial expressions*; after collecting the data from the detected Action Units, a probabilistic histogram table was stored on the Dynamic *Bayesian network*. For visualization purposes, sample lines of this table are illustrated on Table 5.

Concerning the *learning for Bayesian analysis of vocal expressions*; after teaching the system, by pointing which was the correct Vocal Expression for a given input, the result was a histogram table stored on five different files. Each file contained the trained information and its format is summarized on Table 6. These files correspond to the histogram which was the likelihood knowledge for the *Bayesian network*.

6.2 Learning for Decisions

Another part of the memory was devoted to the dynamic rules of the synthesis part. As a result of the learning phase, the responses were taught according to the input V_E , F_E and *SBP*. The result was a histogram that contained the trained information. The format of the histogram trained file is summarized on Table 7. These files corresponded to the histogram which was the likelihood

Table 5 Learning for analysis of facial expressions (α = Action Unit)

	Upper face							
	EB				*CH*		*LE*	
F_E	non	$\alpha 1$	$\alpha 4$	$\alpha 1+4$	non	$\alpha 6$	non	$\alpha 7$
ang	0.01	0.01	0.97	0.01	0.99	0.01	0.01	0.99
fea	0.01	0.01	0.01	0.97	0.99	0.01	0.99	0.01
hap	0.97	0.01	0.01	0.01	0.01	0.99	0.99	0.01
sad	0.01	0.01	0.01	0.97	0.99	0.01	0.99	0.01
neu	0.97	0.01	0.01	0.01	0.99	0.01	0.99	0.01

	Lower face										
	LC			*CB*		*MF*			*MA*		
F_E	non	$\alpha 12$	$\alpha 15$	non	$\alpha 17$	non	$\alpha 20$	$\alpha 23$	non	$\alpha 24$	$\alpha 25$
ang	0.98	0.01	0.01	0.01	0.99	0.01	0.01	0.98	0.01	0.98	0.01
fea	0.98	0.01	0.01	0.99	0.01	0.01	0.98	0.01	0.01	0.01	0.98
hap	0.01	0.98	0.01	0.99	0.01	0.98	0.01	0.01	0.01	0.01	0.98
sad	0.01	0.01	0.98	0.01	0.99	0.98	0.01	0.01	0.98	0.01	0.01
neu	0.98	0.01	0.01	0.99	0.01	0.98	0.01	0.01	0.98	0.01	0.01

Table 6 Since the training files are very big, this table presents just a sample of what the training files contain. This training set is about the learning for vocal expressions

Phrase number	Pitch	Duration	Volume	Vocal expression
5	136.5	3	1494	Neutral
9	159.7	3	1097	Neutral
8	75.1	1	512	Sad
1	110.9	2	4669	Anger

Table 7 Due to the size of the training files, a sample of what the training files contain is presented in this table. This sample is about the learning for synthesis

Phrase number	*V_E*	*F_E*	*SBP*	Response
5	A	A	Sym	r1
9	F	F	Ant	r12
8	H	N	Hum	r4
1	A	N	Sym	r14

knowledge for the first layer (level 0 and level 1) of the *Bayesian network* represented on Fig. 12. The response may vary among the emotional scope, and also across several possible sentences existent in a database for that emotion.

7 Assessments

7.1 Assessment for the Sensory Processing and Feature Extraction

It is known that the classifiers' results depend directly on the effectiveness of the detectors. Usually researchers test the system in optimal conditions and do not do stress tests with different environments. We decided to also define assessments for the used detectors. Once again this was done for both the auditory feature extraction and for the visual (face images) feature extraction. For the sound it was measured using the *standard deviation* (σ_X), *mean* ($\overline{X}$) *and median* ($\widetilde{X}$) where X is a random variable standing for each of our variables (*PT*, *SD* and *VL*) during 100 iterations. Each group of 100 iteration was done in 3 different environments keeping the same phrase, the same performed emotion (neutral) and the same user. The environments were:

1. Good environment: alone with the robot in a room (no background noise);
2. Medium environment: two other people talking normally in the same room with a distance less than 5 meter away (standard noise);
3. Noisy environment: ten persons were asked to talk loudly in the same room (a lot of background noise).

For the images we measured the *percentage of correct face feature extractions* across 100 frames. Ten iterations were done in 3 different environments keeping the same facial expression, the same performed emotion (neutral) and the same user. The environments were:

1. Good environment: person alone with a clean background (no background noise);
2. Medium environment: person alone with a random background (standard noise);
3. Noisy environment: person not alone, other faces were in the image (a lot of background noise).

7.2 Assessment for the Classifiers

7.2.1 Assessment for Automatic Emotion Recognition from Audio Signal

Automatic *emotion recognition* of vocal expressions presented so far in the literature commonly uses features based on: Pitch, the fundamental frequency of the acoustic signal; Energy, also called intensity or volume-level; Speech Rate, the number of words spoken in a time interval or the sentence/phrase duration when the number of the words inside each phrase is known; Pitch contour, the geometrical patterns of the pitch variations; Phonetic features, the pronunciation features. About the classification techniques, we found on the literature: ANNs (Artificial Neural Networks),

HMMs (Hidden Markov Models), Gaussian Mixture density models, Fuzzy membership indexing and maximum-likelihood Bayes classifiers (similar but different from ours). Since there is no common benchmark for such systems, we used a comparison methodology proposed on [29] to show the advantages of our classifier. The following questionnaire was used in Table 10, this questionnaire was defined in [29].

1. Can non professionally spoken input samples be handled?
2. Is the performance independent of variability in subject's sex, physiognomy, age, and ethnicity?
3. Are the auditory features extracted automatically?
4. Are the pitch-related variables utilized?
5. Is the vocal energy (intensity) utilized?
6. Is the speech rate utilized?
7. Are pitch contours utilized?
8. Are phonetic features utilized?
9. Are some other auditory features utilized?
10. Can inaccurate input data be handled?
11. Is the extracted vocal expression information interpreted automatically?
12. How many interpretation categories (labels) have been defined?
13. Are the interpretation labels scored in a context-sensitive manner (application, user, task-profiled manner)?
14. Can multiple interpretation labels be scored at the same time?
15. Are the interpretation labels quantified?
16. Is the input processed in fast or real time?

7.2.2 *Assessment of Automatic Emotion Recognition from Face Images*

Automatic *emotion recognition* from images clearly includes three sub-problems: finding faces, extracting features, and classification. Many of the current systems assume the presence of a face in the scene and do not automatically find faces [30, 31] . However, for example, in [32, 33] a camera was fixed pointing to the human face, so they did not really need to find faces. In HRI area, the camera was always on the robot and not on the human. Most systems assumes good illumination, a clean background and usually they do not provide any automatic or even manual tool to deal with illumination problems. Several improvements have been done in the area of extracting faces [34, 35]. In our case, for finding faces we used the OpenCV haarlike features [35], this method is well known as being independent of illumination problems. Many of the current approaches do not automatically extract the features, do not consider time sequence frames, and it is common that they divide the image in parts instead of analyzing the whole face image at once.

About the classification techniques, we found on the literature: template-based classification [31], fuzzy classification, ANN based classification [30], HMM based classification and Bayesian classification [36, 37]. Since there is no common benchmark for such systems, we will use here a comparison methodology proposed on [29] to show the advantages of our classifier. The following questionnaire was used in Table 11, this questionnaire was defined in [29].

1. Is the input image provided automatically?
2. Is the presence of the face assumed?
3. Is the performance independent of variability in subject's sex, physiognomy, age, and ethnicity?
4. Can variations in lighting be handled?
5. Can rigid head movements be handled?
6. Can distractions like glasses and facial hair be handled?
7. Is the face detected automatically?
8. Are the facial features extracted automatically?
9. Can inaccurate input data be handled?
10. Is the data uncertainty propagated throughout the facial information analysis process?
11. Is the facial expression interpreted automatically?
12. How many interpretations categories (labels) have been defined?
13. Are the interpretation labels user profiled?
14. Can multiple interpretation labels be scored at the same time?
15. Are the interpretation labels quantified?
16. Is the input processed in fast or real time?

7.2.3 *Developed Tools for Assessment of Classifiers*

To be suitable for the comparison methodology proposed in [29], we needed first to measure *the percentage of correct classifications* (hit-rate) for both classifiers (audio and video). This was done by comparing the classified result to the expected result. We created graphical interfaces (see Figs. 16 and 17) to help accomplish this task. On those interfaces, for both audio and video; the system tester could see the result of classifications on real time, and click on what he/she expected as the result. When the system tester clicked on the expected result; a benchmark routine saved both the real time classification and the expected expression in a file, for further statistical calculation of *the percentage of correct classifications* (hit-rate).

7.3 Assessments for Synthesis: Study Case of Automatic Humor Generation

It is difficult to measure how funny (henceforth meaning humorous) a system can be during Social Human Robot Interaction. On the literature [38, 39], specially those from the European project called "Hahacronym", we found descriptions of results but no detailed descriptions of assessments.

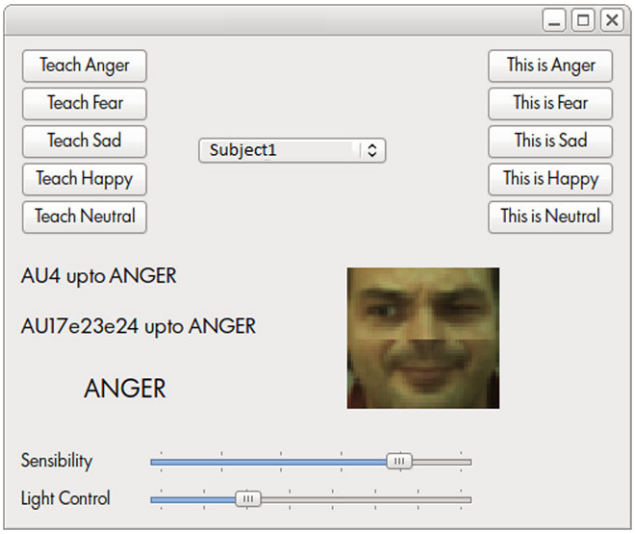

Fig. 16 Graphical interface for face classifier and assessment

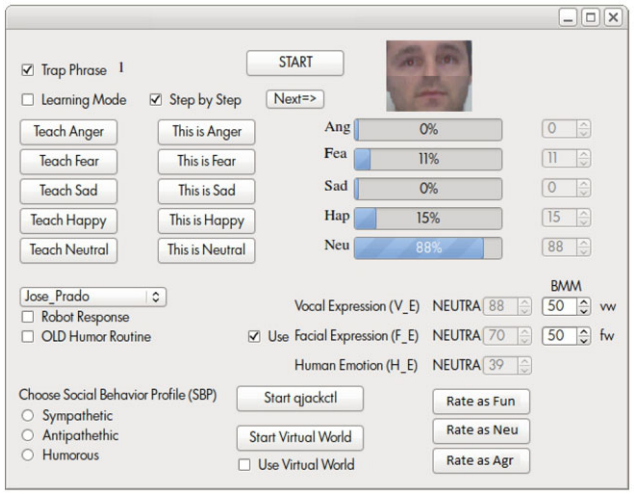

Fig. 17 Our graphical user interface: it is possible to trap the phrase in a loop, put the system in learning mode in order to fill-out the *Bayesian network*. To go step-by-step or to release the system to go "fast". The buttons: "This is *anger*" and so on are used to manually score if the classification was correct or not. There are also possibilities to choose the *SBP* among Sympathetic, Antipathetic and Humorous. This interface also allows to enable or disable our Virtual World which will pop-up in another window if selected. Also if the user goes step-by-step it is possible to rate the robot's response as Funny, Neutral or Aggressive

However it is understandable that they did their experiments with several people, while an external agent did a manual classification of how happy the person was with the performance of that system. In [40], the description of assessments were more clear when the system was shown to children, and what was considered a joke was also manually measured. They followed an assessment protocol for measuring the "jokiness" of each response proposed on [41]. Previously on [41] it was measured the average "jokiness", "funniness" and "heard before" scores for each text, with their set number and source. Scores for "jokiness" range from 0 (none of the children who were asked to rate the text thought it was a joke) to 1 (all of the children who were asked to rate the text thought it was a joke). Scores for "funniness" ranged from 1 to 5, with 1 meaning "not funny at all" and 5 meaning "very funny". Scores for "heard before" ranged from 0 (none of the children who were asked to rate the text had heard it before) to 1 (all of the children who were asked to rate the text had heard it before).

7.3.1 How We Do Assessments of Automatic Humor Generation

Considering the state-of-art, there is no common benchmark for this type of system. There are existent ideas for assessments [29], which we reinterpreted by defining our own assessments.

Our system can be set to three different *SBP* {Sympathetic, Antipathetic and Humorous}. The *Bayesian network* was initially trained in order to contain response scopes that match with the expected *social behavior profile*, and randomness was added to the decision process of response. To measure that match, we defined the assessment protocol as being: after the system is trained, the subject interacts with the system during a story board of 9 phrases from the robot and 9 phrases from the human. After each phrase, some seconds are given for the human to perform a facial expression. Thus, the *BMM* takes place by giving H_E. Later that H_E is merged with *SBP* and the robot acts by speaking a response phrase chosen randomly among the 3 most probable answers given by the inference. This response phrase was then rated by the subject as being exclusively: funny (F), neutral (N) or aggressive (A).

The robotic *SBP* was then changed and the differences on the evaluation were collected. In Fig. 17 is our implemented graphical interface that helped us during the process of evaluating our structure.

8 Experiments

8.1 Platform Setup

Initially, our robotic platform [42] was designed in a Segway base and a robotic head with 4 degrees of freedom. Later it migrated to a Scout platform and a head with 2 degrees of freedom and a support for a screen was added to show the expressions. Furthermore a retro-projectable mask was built for a better interactive interface (see Figs. 18 and 19(a)).

In both cases the robotic technology used as the experimental platform had an active vision system. This feature allowed the robot to move its head towards the tracked face

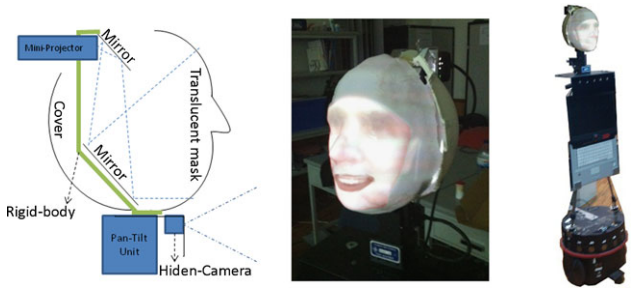

Fig. 18 Retro-projected translucent mask and the robot assembled

Fig. 19 (**a**) One version of our robot: Scout based platform and a head with 2 degrees of freedom. (**b**) Our virtual world is another option of interaction instead of the robot. A real person can look to the camera, and speak at the microphone; where an avatar mimics this person and the other avatar simulates the robot

before starting to move its body, thus, the robot avoided unnecessary movements with the body structure.

Furthermore, as another platform; we developed a 3D virtual world as a "Blender game", where the same core of interactions could be used both over the real robot and/or inside the virtual world; see Fig. 19(b). We generated 14 meshes of heads from 14 persons of our lab, so that we used the face of the real person on the avatar that mimics the person. Stereo vision systems were also an option that we tested but did not pursue. According to [43], the background segmentation allowed by the stereo vision can be used to improve the selection of which user to interact with.

8.2 Experiments on the Sensory Processing

8.2.1 Results for Auditory Sensory Processing and Discussion

According to what was defined for our assessments in Sect. 7.1, the results were collected and can be seen in Table 8. As expected, the standard deviation in a good environment was acceptable, because even the same person when he/she repeats the same sentence with the same vocal expression does not produce exactly the same audio signal. On the medium environment, the standard deviation increased very little; the mean and median were quite similar to the good environment since we were using a microphone close to the mouth of the user, the influence in medium environment did not significantly affect the results of the sensory processing. However in the noisy environment the standard deviation increased a lot, specially for *SD*. That's because our phrase ends automatically with a silence detector implemented based on the signals' amplitude, this was highly disturbed by the noisy environment. Also notice that the mean and median of Energy (*VL*) significantly increased in the noisy environment. In noisy environment, the pitch values seemed to be completely wrong for some cases because the noisy signal may be composed by peaks of frequency. Thus, we concluded that we can use our system only in good or medium environments. Thus, all experiments over our classifiers were done in a "medium environment" for the feature extractors.

Table 8 GE stands for Good Environment, ME stands for Medium Environment and NE stands for Noisy Environment. Standard deviation, mean and median of *PT*, *SD* and *VL* are shown, respectively, over the three different environments as defined on the assessments. One hundred tests were done for each environment, keeping phrase and user the same

	$abs(PT)$			SD			VL		
	σ_{PT}	$\overline{PT}$	$\widetilde{PT}$	σ_{SD}	$\overline{SD}$	$\widetilde{SD}$	σ_{VL}	$\overline{VL}$	$\widetilde{VL}$
GE	10.7	124.5	133	0	3	3	105.1	1613.9	1590
ME	15.4	126.2	133	0.3	3.10	3	153.1	1623	1590
NE	103.2	181.7	139	1.77	4.36	3	235.4	1777.5	1707

Table 9 GE stands for Good Environment, ME stands for Medium Environment and NE stands for Noisy Environment. The percentage of correct face feature extractions are shown, respectively, over the three different environments as defined on assessments. One hundred frames were collected for each environment

	GE	ME	NE
Correct %	98%	78%	63%

8.2.2 Results for Visual Sensory Processing and Discussion

According to what was defined for our assessments in Sect. 7.1, the results were collected and can be seen in Table 9. We realized that the facial detector we were using allowed us to work only in the good environment. The percentages shown in Table 9 are not satisfactory for the medium and noisy environments.

8.3 Experiments on the Classifiers

8.3.1 Bench Mark over the Vocal Expression Emotion Classifier

At first, the phrase was blocked so that the same phrase was repeated several times with different intonations. This procedure was done during the learning phase, when the user repeated the same phrase 50 times. From these 50 audio files, the features were extracted and this set of features was kept as the *trained set* for the current phrase. This *trained set* belongs to the user who trained it and was used for that user. A short dialog containing 9 phrases was used to guide the experimental tests, and, thus, the training procedure was repeated for each of the phrases used during the conversation. Therefore, a total of 450 sentences were used as the *trained set*.

After the learning phase, 129 sentences were tested into four batteries of tests. Nine in the first battery of tests, twenty nine in the second, forty one in the third, and fifty in the fourth battery of tests. These tests were done over the Vocal Expression Classifier; the results can be seen in Fig. 20.

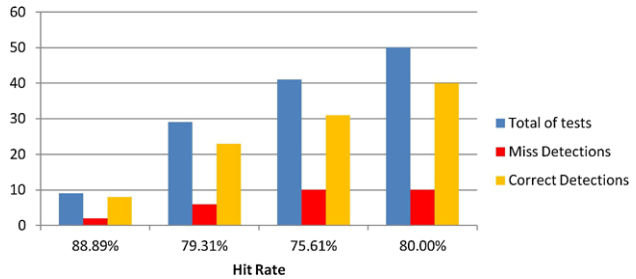

Fig. 20 Four batteries of tests were done with clicking on the expected classification while saving the current classification. All the vocal expressions were randomly mixed during the tests while a person was speaking all the five considered possible vocal expressions. The average percentage of correct classifications is 80.92%

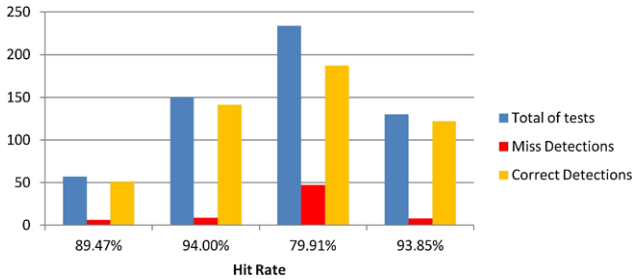

Fig. 21 Four batteries of tests were done with clicking on the expected classification while saving the current classification. All the Action Units were randomly mixed during the tests while a person was performing all the five considered facial expressions. The average percentage of correct classifications was 89.27%

8.3.2 Bench Mark over the Facial Expression Emotion Classifier

According to what was defined on the assessments, after the system was trained for the possible facial expressions, we did a battery of tests over the Facial Expression Classifier; the results can be seen in Fig. 21.

Notice that our AU Detector is also dependent on the efficiency of the OpenCV haar-like features face detector, nevertheless, the results were satisfactory.

8.4 Comparison of Classifiers with State-of-Art and Discussion

8.4.1 Audio Recognition Comparison

By using the comparison methodology proposed on [29], Table 10 shows the properties of state-of-art systems and also our system. Other methods claimed to achieve a higher percentage of correct classifications, however they were not fully real-time, usually they used a database input of utterances instead of capturing directly from the microphone. Our approach captured directly from the microphone, and did every calculation in less than 1 second, being thus useful in our proposed structure and suitable for HRI applications.

Table 10 Properties of state of art approaches to automatic *emotion recognition* from audio signals

Reference	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Test results
Dimitrius'06 [44]	×	U	•	•	•	•	•	•	×	×	•	U	×	•	•	×	Correct from 58.6% to 94.4% depending on the applied method
Nicolaou'10 [33]	×	T	•	•	•	U	U	U	•	×	•	3	×	×	×	U	Correct from 61.19% to 91.96% depending on the method and the expression
Our proposed approach	×	T	•	•	•	•	×	×	•	×	•	5	×	•	•	•	Correct 81%, mean of all expressions

Legend: •= "yes", × = "no", U = unknown, T = handle speech samples of (known) subjects on which it has been trained

Table 11 Properties of state of art approaches to automatic *emotion recognition* from facial images

Reference	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Test results
Cohen'02 [36]	•	•	U	U	×	U	×	•	×	×	•	7	×	×	×	×	12600 frames, 5 subjects, Correct: 65%
Wuhan'04 [31]	×	•	×	×	×	×	×	•	×	×	•	7	×	×	×	×	213 frames, 10 women, Correct 77%
Nicolaou'10 [33]	×	•	U	×	•	×	•	•	×	•	•	6	×	•	•	×	Correct 91.76%
Pantic'09 [32]	•	•	•	×	•	×	•	•	×	•	•	6	×	•	•	•	Correct from 61% to 93% depending on the expression
Gayatri'08 [30]	×	•	•	×	×	×	×	×	×	×	•	7	×	×	×	×	U frames, 40 subjects, 94.73%
Our proposed approach	•	×	•	×	•	×	•	•	×	•	•	5	×	•	•	•	560 frames, 1 subject, Correct 89.27%

Legend: • = "yes", × = "no", U = unknown, T = handle images of subjects on which it has been trained

8.4.2 Visual Recognition Comparison

By using the comparison methodology proposed on [29] and the percentage of correct classifications acquired by our tests, Table 11 shows the properties of some of the current existent systems for the same purpose and we included our system on this table for a fair evaluation.

It is important to use this table comparison, because for example in [45] and [30], they claim to achieve a high percentage of correct classifications, however features were not extracted automatically. In [30], Gimp software distance measurement was used to manually extract the features and the Cohn-Kanade database [46] was used, so, in [30], lighting variations also cannot be handled and neither was it necessary to detect faces because all the images were already faces. In Wuhan'04 [31], the two eye pupils needed to be selected manually, no face feature extraction was done, JAFFE [47] database was used, and the system was not real time. A lot of progress was done in [32, 33], and they are quite likely the state of art in this area. Our classifier is a simpler version, however, it has some advantages in some aspects that allows it to be well suited for our proposed structure. For example, we did not use a database of faces, we captured video stream from the robot camera and thus we did not assume that a face was there. We could deal with rigid head movements because the face feature extraction was done statically. Our system was fully automatic and it ran in real time. The percentage of correct classifications of our facial expression classifier was acceptable and suitable for HRI applications.

Table 12 With *SBP* set to Antipathetic (**a**), Sympathetic (**b**) and Humorous (**c**); this table contains the subject's evaluation about the robot's response to each sentence of the story board. According to the defined assessments: F was given if the human felt the robot's response was funny, N to neutral and A to aggressive

(a)

Robot resp.	r1	r2	r3	r4	r5	r6	r7	r8	r9
Subject 1 eval.	N	A	N	A	A	A	A	N	N
Subject 2 eval.	A	A	A	A	N	A	A	A	A
Subject 3 eval.	A	A	A	A	A	A	A	A	A

(b)

Robot resp.	r1	r2	r3	r4	r5	r6	r7	r8	r9
Subject 1 eval.	F	N	N	N	N	N	N	N	N
Subject 2 eval.	N	N	N	N	N	N	N	A	N
Subject 3 eval.	F	N	N	N	F	N	N	N	F

(c)

Robot resp.	r1	r2	r3	r4	r5	r6	r7	r8	r9
Subject 1 eval.	F	F	F	N	N	F	F	N	F
Subject 2 eval.	N	N	F	F	F	N	F	N	F
Subject 3 eval.	F	F	F	F	F	F	F	F	F

8.5 Experiment on the Synthesis and Fusion with *SBP*

After trained, the system was presented to 3 male subjects where they rated the robot response among the scope defined in assessments (Sect. 7). Ten experiments and evaluations were done per subject and we obtained the Tables 12(a), 12(b) and 12(c).

9 Conclusions and Future Work

Our proposed methodology to support the interaction between human and robot is a novel approach. It constructs in the robot a model of emotive responses that is similar to the premises established for humans. Moreover, it synthesizes facial expressions and vocalization with lips synchronization, based on the inferred emotive response. The strategy is based on neuropsychology and our model is expandable to more modalities. It is a contribution in the direction of having robots with automatic emotional response. It was defined a set of assessments and then these assessments were used over the experiments to show the performance of our algorithm in comparison with the state of the art. There are several new contributions in this study, we consider that the key contributions are the real time classifiers from both audio and video. One of the main limitations is the learning; it is still necessary to train the system before using it, this costs an extra time for adequate training when implementing it to real world applications. From the results achieved, we can conclude that the expression analysis, both vocal and facial, classify and converge as expected. Additionally, the proposed fusion included in the synthesis enhances the social quality of the interaction. As future work we want to further explore the possibilities of the fusion, specially the *humorous social behavior profile*.

References

1. Gratch J, Marsella S, Petta P (2008) Modeling the cognitive antecedents and consequents of emotion. Cogn Syst 10(1):1–5
2. Kidd CD, Breazeal C (2007) A robotic weight loss coach. In: Proceedings of the twenty-second conference on artificial intelligence, Menlo Park, CA. AAAI Press, Menlo Park
3. Schroder M (2010) The semaine api: Towards a standards-based framework for building emotion-oriented systems. Adv Hum Comput Interact 2010:319406. doi:10.1155/2010/319406/2010/319406. 21 pp.
4. Lee CM, Narayanan SS, Pieraccini R (2002) Classifying emotions in human-machine spoken dialogs. In: ICME
5. Wang Y, Guan L (2005) Recognizing human emotion from audiovisual information. In: ICASSP IEEE
6. Cowie R, Douglas-Cowie E, Karpouszis K, Caridakis G, Wallace M, Kollias S (2007) Recognition of emotional states in natural human-computer interaction. School of Psychology, Queen's University
7. Darwin CR (1872) The expression of the emotions in man and animals, 1st edn. Murray, London
8. Ekman P, Friesen WV, Hager JC (2002) Facial action coding system—the manual. A human face
9. Ekman P, Friesen W (2003) Unmasking the face: A guide to recognizing emotions from facial clues. Malor Books, Cambridge
10. Ekman P, Rosenberg E (2004) What the face reveals: basic and applied studies of spontaneous expression using the facial action coding system (FACS), 2nd edn. Oxford University Press, London
11. Damasio A (2003) Looking for Spinoza. Harcourt Brace & Co., San Diego. ISBN:978-0-15-100557-4
12. Damasio A (2000) The Feeling of what happens. Harcourt Brace & Co., San Diego. ISBN:978-0-15-601075-7
13. Spinoza B (1677) Ethics
14. Chaitin GJ (2010) Meta math!: the quest for omega. Pantheon, New York. The University of Michigan
15. Lori N, Blin A (2010) Application of quantum Darwinism to cosmic inflation: an example of the limits imposed in Aristotelian logic by information-based approach to Godel's incompleteness. Found Sci 15:199–211
16. Lori NF, Jesus P (2010) Matter and selfhood in Kant's physics: a contemporary reappraisal. In: Balsemão Pires E, Nonnenmacher B, Büttner-von Stülpnagel S (eds) Relations of the self. Imprensa da Universidade de Coimbra, Coimbra, pp 207–226
17. Levine PA (1997) Waking the tiger—healing trauma. North Atlantic Books, Berkeley
18. Damasio A (2010) Self comes to mind: constructing the conscious brain. Pantheon, New York
19. Evers K (2009) The empathetic xenophobe: a neurophilosophical view on the self. In: Centre for research ethics and bioethics, (CRB), Uppsala University. The text is adapted from Chap. 3 in Evers (2009): Neuroethique. Quand la matiere s eveille, Editions Odile Jacob, Paris, and was originally presented in an earlier version at College de France, Paris, 2006
20. George S, Leroux P (2002) An approach to automatic analysis of learners social behavior during computer-mediated synchronous conversations. In: Cerri S, Gouarderes G, Paraguacu F (eds) Intelligent tutoring systems. Lecture notes in computer science, vol 2363. Springer, Berlin, pp 630–640 [Online]. Available: doi:10.1007/3-540-47987-2_64
21. Kau AS, Tierney E, Bukelis I, Stump MH, Kates WR, Trescher WH, Kaufmann WE (2004) Social behaviour profile in young males with fragile x synfrome: characteristics and specificity. Am J Med Genet 126:9–17
22. Dahlbäck N, Jönsson A, Ahrenberg L (1993) Wizard of oz studies: Why and how. In: Proceedings of the international workshop on intelligent user interfaces, Orlando, FL. ACM, New York, pp 193–200
23. Klemmer S, Sinha A, Chen J, Landay J, Aboobaker N, Wang A (2000) Suede: a wizard of oz prototyping tool for speech user interfaces. In: CHI letters: Proceedings of the ACM symposium on user interface software and technology, vol 2, pp 1–10
24. Ernst M, Bülthoff H (2004) Merging the senses into a robust percept. Trends Cogn Sci 8(4):162–169
25. Sondhi M (1968) New methods of pitch extraction. IEEE Trans Audio Electroacoust 16:262–266
26. Boersma P, Weenink D, Eletronic University of Amsterdam [Online]. Available: http://www.fon.hum.uva.nl/praat/
27. Invertions S (2010) Eletronic [Online]. Available: www.facegen.com
28. Intel (2006) Intel open source computer vision library, http://www.intel.com/technology/computing/opencv
29. Pantic M, Rothkrantz LJM (2003) Toward an affect-sensitive multimodal human-computer interaction. Proc IEEE 91(9):1370–1390
30. Paknikar G (2008) Facial image based expression classification system using committee neural networks. PhD dissertation, The Graduate Faculty of The University of Akron
31. Wuhan (2004) Facial expression recognition based on local binary patterns and coarse-to-fine classification. In: Fourth international conference on computer and information technology (CIT'04), vol 16
32. Pantic M (2009) Facial expression recognition. In: Encyclopedia of biometrics, pp 400–406
33. Nicolaou MA, Gunes H, Pantic M (2010) Audio-visual classification and fusion of spontaneous affective data in likelihood space. In: ICPR, pp 3695–3699
34. Yang MH, Kriegman DJ, Ahuja N (2002) Detecting faces in images: a survey. IEEE Trans Pattern Anal Mach Intell 24:34–58

35. Viola P, Jones MJ (2001) Rapid object detection using a boosted cascade of simple features. In: IEEE CVPR
36. Cohen I, Sebe N, Garg A, Lew M, Huang T (2002) Facial expression recognition from video sequences. In: Proc ICME, pp 121–124
37. Sebe N, Lew M, Cohen I, Garg A, Huang T (2002) Emotion recognition using a Cauchy naive Bayes classifier. In: Proc ICPR, vol 1, pp 17–20
38. Stock O, Strapparava C (2003) Getting serious about the development of computational humor. In: Proceedings of the 8th international joint conference on artificial intelligence (IJCAI), pp 59–64
39. Stock O, Strapparava C (2005) The act of creating humorous acronyms. J Appl Artif Intell 19:137–151
40. Ritchie G (1998) Prospects for computational humor. In: Proceedings of 7th IEEE international workshop on robot and human communication, pp 283–291
41. Binsted K Pain H, Ritchie G (1997) Children's evaluation of computer-generated punning riddles. Department of Artificial Intelligence, University of Edinburgh
42. Prado J, Lobo J, Dias J (2010) Sophie: social robotic platform for human interactive experimentation. In: 4th international conference on cognitive systems, COGSYS 2010, ETH Zurich, Switzerland
43. Prado J, Santos L, Dias J (2009) Horopter based dynamic background segmentation applied to an interactive mobile robot. In: 14th international conference on advanced robotics, ICAR09, Munich, Germany
44. Ververidis D, Kotropoulos C (2006) Emotional speech recognition: resources, features, and methods. Speech Commun 48(9):1162–1181 [Online]. Available: http://www.sciencedirect.com/science/article/B6V1C-4K1HCKM-1/2/3c1a10a68e9fe662b07918424294495a
45. Lyons M, Budynck J, Akamatsu S (1999) Automatic classification of single facial images. IEEE Trans Pattern Anal Mach Intell 21:1357–1362
46. Kanade T, Cohn V, Tian Y, (2000) Cohn-Kanade au-coded facial expression database [Online]. Available: http://vasc.ri.cmu.edu/idb/html/face/facial_expression/
47. Kamachi M, Lyons M, Gyoba J (1998) The Japanese female facial expression (jaffe) database [Online]. Available: http://www.kasrl.org/jaffe.html

José Augusto Prado born in Rio de Janeiro RJ, southeast of Brazil. He completed his Bachelor degree in Computer Science at 2002. His final project of graduation was a beneficent database software which was donated to Fundação da Criança Renal. He finished his M.Sc. degree in Computer Science and Probabilistic Algorithms at the Federal University of Paraná in November 2005. During his Ph.D. he worked as a fellow within the European Project BACS (Contract no. FP6-IST-027140) as a researcher. Currently he is a Ph.D. student sponsored by a scholarship from the Portuguese Foundation for Technology and Sciences (SFRH/BD/60954/2009), at the Institute of Systems and Robotics, in the Faculty of Sciences and Technology, University of Coimbra, Portugal.

Carlos Simplício born in Vila Real de Santo António, Portugal. He completed his licentiate degree on Electrical Engineering, specialisation in Industrial Systems, at University of Coimbra in 1990. He finished his M.Sc. in Systems and Technologies of Information, specialisation in Control Systems, at University of Coimbra in 2003. To obtain the master degree, he developed a "System to Identify Trafic Signs, based on Projective Invariants". Currently he is a Ph.D. student at the Institute of Systems and Robotics, in the Faculty of Sciences and Technology, University of Coimbra. From 1994, he teaches robotics, computer vision and electronics in the School of Technology and Management, Polytechnic Institute of Leiria.

Nicolás F. Lori born in Buenos Aires, Argentina; has a Ph.D. degree on Physics at Washington University in Saint Louis, USA, specialization in Diffusion Tensor Tracking of Neuronal Fiber Pathways in the Living Human Brain, May 2001. Nicolás Lori conducts his research activities at the Institute of Biomedical Research in Light and Image (IBILI), Faculty of Medicine, University of Coimbra. Nicolás Lori's research areas are diffusion MRI, mathematical modeling of joint neuronal dynamics (both electronic and biologic); with activities and contributions on the first field since 1999 and on the second since 1996. He has several publications on Scientific Reports, Conferences, Journals and Book Chapters. Nicolás Lori teaches biomedical engineering courses at the Physics Department, Faculty of Science and Technology, University of Coimbra. He has been responsible for courses on biostatics, and MRI. He is also responsible for the supervision of graduate students on the field of biomedical engineering.

Jorge Dias born in Coimbra, Portugal and has a Ph.D. degree on Electrical Engineering at University of Coimbra, specialisation in Control and Instrumentation, November 1994. Jorge Dias conducts his research activities at the Institute of Systems and Robotics (ISR-Instituto de Sistemas e Robótica) at University of Coimbra. Jorge Dias' research area is Computer Vision and Robotics, with activities and contributions on the field since 1984. He has several publications on Scientific Reports, Conferences, Journals and Book Chapters. Jorge Dias teaches several engineering courses at the Electrical Engineering and Computer Science Department, Faculty of Science and Technology, University of Coimbra. He is responsible for courses on Computer Vision, Robotics, Industrial Automation, Microprocessors and Digital Systems. He is also responsible for the supervision of Master and Ph.D. students on the field of Computer Vision and Robotics.