# An Affective Robot Guide to Museums

**DimitriosVogiatzis[1], Constantine D. Spyropoulos[1], Stasinos Konstantopoulos[1], Vangelis Karkaletsis[1], Zerrin Kasap[2], Colin Matheson[3], Olivier Deroo[4]**

**Abstract.**

The basic goal of human robot interaction is to establish an effective communication between the two parties. In particular, robot emotion, speech, and facial expressions determine the way humans regard the robot, and they are deemed as essential for a natural form of communication. Addressing those issues is the focal point of this paper, while as a testbed we deployed a robot in a museum, where it serves as a guide to visitors. Our aim is a system that exhibits basic rational and intelligent behaviour.

## 1 INTRODUCTION

In this work, we focus on the affective module of the robot platform that is being developed in the context of the INDIGO project (http://www.ics.forth.gr/indigo/). The INDIGO robot platform (depicted in Figure 1) is able to interact with humans in natural language. Thus, it includes speech recognition and synthesis (TTS) functionalities, an advanced dialogue system, a natural language generation engine, and a touch screen which can be used to control the interaction. As the robot is mobile, it is able to navigate in crowded areas and to recognise basic human gestures. Finally, it is equipped with an animatronic head on which different facial characteristics can be exhibited. The robot is designed to be responsive to the desires of the human visitors, allowing them to influence museum tours and ask simple questions.

The test case of INDIGO is the hall of the dome of the Foundation of the Hellenic World (http://www.tholos254.gr/projects/agora/en). INDIGO will exhibit natural and intelligent behaviour as a result of the *affective* aspects that are incorporated.

In the next sections, we focus on how the robot detects, represents, generates and expresses emotions. In particular, in section 2, we describe the INDIGO architecture, in section 3 we present how robotic emotions are enacted, in section 4 we describe how robotic emotions are communicated to the visitor and section 5 presents conclusions and directions for future work.

## 2 INDIGO ARCHITECTURE

The overall INDIGO architecture appears in Figure 2. Knowledge structures are depicted as rounded boxes and processes as square boxes. Briefly, visitor's actions (utterances and gestures) are analysed by an array of language and vision tools. The results of the analysis are fused and passed to the *dialogue and action manager (DAM)*, which is responsible for generating a robot action in response. The results of the linguistic analysis are also passed to the robot personality module, which calculates an emotional appraisal of the visitor's



**Figure 1.** INDIGO robot.

utterance. If the robot's action is to be the description of another object, the DAM consults the robot personality module to receive affect parameters associated with the possible object to describe (for instance, whether the robot enjoys talking about this particular class of objects).

These parameters are used by the DAM to make the final decision about the object to describe, and this choice also generates an emotional appraisal. Both user-action and robot-action appraisals are forwarded to the *emotional state machine*, which updates the *mood* and the *emotional state* of the robot and results in an annotation of the robot action. This annotation is used by TTS and by the animatronic head to modulate the voice and facial expression of the robot. The knowledge structures are authored using ELEON, a tool developed at NCSR "Demokritos" (http://www.iit.demokritos.gr/~eleon). The authoring language is OWL-DL for the domain ontology, RDF for the language resources associated with the ontology (lexicon, grammar), and RDF for the user models, which describe static user preferences (robotic models are analysed in the next section).

### 2.1. Dialogue and Action Manager

The DAM is implemented using the TrindiKit, which is based on the *Information-State (IS)* and *Update* model of dialogue [4]. In this model the IS is used to store information on 'the current state of the dialogue' in a very broad sense, and a series of
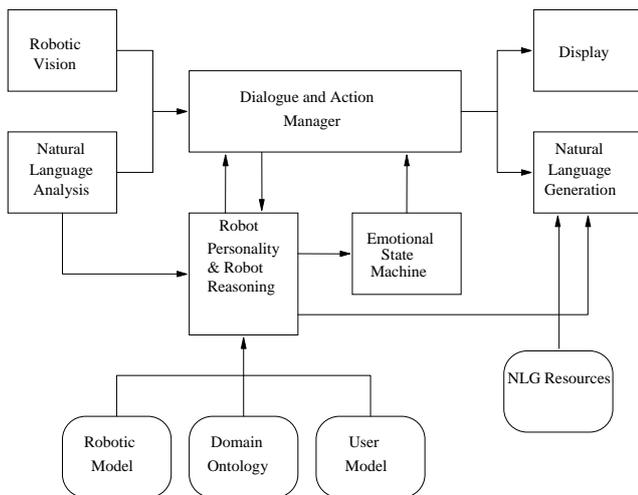
**Figure 2.** INDIGO system architecture.

$$
IS: \begin{bmatrix} OBJECTS: \begin{bmatrix} VISITED: \{o5, o32, o21\} \\ POSITION: o5 \end{bmatrix} \\ DH: [direct(sys,follow),ask(usr,see,stoa), \ldots] \\ LU: \begin{bmatrix} MOVES: direct(sys,follow) \\ SPEAKER: sys \end{bmatrix} \end{bmatrix}
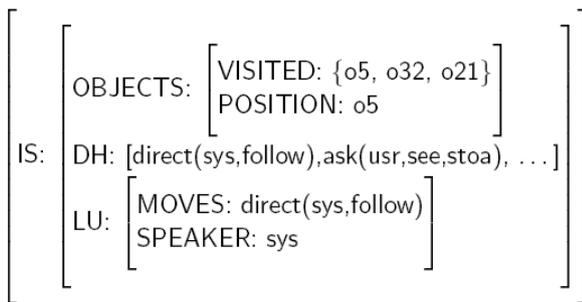$$

**Figure 3.** Part of an INDIGO Information State.

```
rule( assertNoMore,
    [
      is/lu/speaker == usr,
      in( is/lu/moves,  utt(no) ),
      fst(  is/dh, act(sys, ask, more ))
    ],
    [
      push( /is/dh, act(usr, assert, no_more) ),
      push( /is/dh, act(sys,goodbye) )

    ] ).
```

**Figure 4.** An example update rule.

update rules determine how the user's input affects the IS. An example of part of an IS is shown in Figure 3 as an attribute-value matrix, and a TrindiKit rule is provided in Figure 4.

The IS in Figure 3 has an OBJECTS element which records the current position and any previously-visited exhibits, and it also records schematic representations of speech acts as an ordered list in the dialogue history (DH) element. In this case the last two acts were a request from the user to see a stoa, and a consequent directive from the robot to follow:

> USR: Can I see a stoa?

> SYS: Follow me please.

The latest utterance (LU) element contains a representation of the most recent utterance (or action). In this case, of course, the system has just said "follow me".

Update rules are usually simple sets of conditions and effects. In the example in Figure 4 there are three conditions – the last speaker was the user, the utterance contents were "no", and the first thing on the DH list represents the system asking "do you want to hear more?". The two effects are that the user's utterance is interpreted as the assertion "I don't want to hear any more", which is recorded in the DH along with the system's "goodbye" response (the dialogue will end at this point).

In INDIGO the input to the DAM is a combination of information from vision, speech and the touch screen, all of which are fused into a single representation in XML format. The output, also in XML, is multimodal in that it can contain instructions to move the robot platform and/or the robotic head as well as the input to the speech output and emotion modelling modules.

## 2.2. Automatic Speech Recognition

Part of the Natural Language Analysis module of INDIGO is the automatic speech recognition (ASR) module. It supports relatively simple user utterances in the context of system-initiated and task-specific human-robot dialogues. This task is already a challenge because ASR systems are able to perform accurately using closed-talk microphones in non noisy environments, but suffer from large word accuracy degradation when they are used in potentially noisy, densely-populated environments using far talk microphones –as it is the case in the premises of the Foundation of the Hellenic World. Robustness in such difficult environments has not been investigated up to now. The strategy chosen in order to increase the robustness of the ASR system to the museum noises is twofold:

- adaptation of the acoustic models of the speech recognition system in order to cope with classical background noise.
- use of noise robust techniques like Multi-band approach [11,12].

In addition, multiple far talk microphones will be tested in order to obtain the best performances regarding ASR robustness. The final system will be tested on recorded data in the museum with real visitors talking with the robot.

## 2.3. Natural Language Generation Engine

Natural OWL is a natural language generation engine mostly geared towards generating natural language descriptions of ontological entities [2]. Natural OWL generates descriptions that are dynamically customized to the current visitor. Also, comparisons are possible between the entities and the classes of the ontology.

# 3 ENACTMENT OF ROBOT EMOTION

The robot's affective component can be described in terms of three entities: the *personality,* the *emotional state,* and the *mood.* These are discussed below.

## 3.1. Personality

Emotional variation is achieved through synthetic personality models that estimate the emotional appraisal of each dialogue act but also parameterize various INDIGO components in order to externalize emotional and other aspects of the robot personality.

The robot personality is grounded on *profile attributes* that represent subjective views of the entities in the domain ontology. As an example, consider the *interest level* of entities, which is a profile-specific numerical attribute of the individuals, properties, and classes of the domain ontology. Individual and class interest levels represent how interested a robot is in describing a particular exhibit or class of exhibits; Property interest levels control which facts known about an exhibit will be used in order to describe it. So, for example, one profile might favour using architectural order to describe a building whereas another might prefer to provide historical facts.

These profiles can be associated either with individual robots or with user stereotypes. In the former case they are interpreted as robotic personality and individuality attributes. In the latter, they parameterize the interaction to specific audience types, such as *child, expert,* or *general public.*

Profiles are manually authored with the aid of the ELEON authoring tool [6], and work is underway to extend ELEON so that the authoring process is supported by an intelligent back-end which predicts profile attribute values based on already provided values and the ontological relations known for the domain.

As a very simple example, class interest propagates to the class' instances unless explicitly overridden. More complex inference involves property values "flowing" along the various ontological relations, so that, for example, individuals with interesting properties are also predicted to be interesting themselves.

At the core of robot personality is the process that combines the various ground attributes into the parameters required by the various interaction components [5]. In the approach chosen in INDIGO, profile attributes are raised to *many-valued classes,* where class membership is not a binary valuation but a degree in the [0,1] range. User interest, robot interest, and a pre-defined agenda are used to assert that domain ontology entities belong to various such classes. In the case of interest levels, for example, the following classes are defined: RobotInteresting and UserInteresting based on robot profile and user stereotype, and AgendaInteresting based on the pre-defined museum tour.

The combination rules are a many-valued logic programme that pragmatically interprets the OCEAN personality model [1]. OCEAN defines static personality traits as five parameters:

1. Openness (O): openness to experience new things, being imaginative, intelligent and creative.

2. Conscientiousness (C): indicates responsibility, reliability and tidiness. Conscientious people think about all the outputs of their behaviors before taking action and take responsibility.

3. Extroversion (E): Outgoing, sociable, assertive and energetic to achieve his/her goals.

4. Agreeableness (A): trustable, kind and cooperative considering other people's goals and is ready to surrender his/her own goals.

5. Neuroticism (N): anxious, nervous, and prone to depression, lack of emotional stability.

Each robot personality defines the class of Interesting exhibits in a different way, depending on the OCEAN personality of the robot.

Therefore, for example, consciousness involves AgendaInteresting in the definition of Interesting, openness means UserInteresting is taken into account, and extroversion influences membership in DescrAtLength, the class of exhibits that receive lengthy descriptions, and so on [5]. These inferred membership degrees are then used to parametrize the DAM (choosing what to describe next), the Emotional State Machine (providing emotional appraisal of the next robot action) and NLG (description length) (see also Figure 1).

## 3.2. Emotional state and Mood

Dialogue actions—user as well as robot actions—have an impact on the mood and emotional state of the robot, which, in turn, are used to drive speech synthesis and facial expressions. This impact is represented using the OCC model, which defines 22 elementary emotions, such as joy, gratitude, pride, distress, disappointment, and so on. In the OCC model, agent's concerns in an environment are divided into goals (desired states of the world), standards (ideas about how people should act) and preferences (likes and dislikes). In a later work [4], Ortony himself found the 22 distinct emotion types too complex for the simulation of believable characters and decreased the number of labels to 12, 6 being positive (joy, hope, relief, pride, gratitude, love) and 6 being negative (distress, fear, disappointment, remorse, anger). We also use these 12 emotion labels in our emotional model.

In INDIGO, OCC vectors are generated for both user and robot dialogue acts. User actions are linguistically analysed to extract the impact of the manner of the user's utterance. That is, polite or impolite linguistic markers trigger the appropriate OCC appraisal. The appraisal of robot actions, on the other hand, reflects the robot's eagerness or aversion to fulfilling a user request, depending on the robot personality model and the interest parameters of the requested content. The appraisal vectors are used by an emotional-state machine which updates the robot's emotional state to reflect the latest emotional appraisal received, taking into account the OCEAN personality traits and the mood of the robot. Moods and emotions can be differentiated based on three criteria: *temporal, expression* and *cause* [7]. Moods last longer than emotions and they are not associated with a specific event. In other words, emotions modulate actions while moods modulate cognition. The relation between mood and emotions is two-way. Mood affects the

appraisal of events and decides which emotion will be triggered and with what intensity. For example, if the person is in an anxious mood, he/she will be easily disappointed by bad events and with higher intensity. Emotions can also cause a particular mood to occur. For example, a person in a bored mood can change to a more positive mood after some positive emotional appraisals from the environment. Moods are usually represented with continuous dimensions rather than being discrete labels. We use Mehrabian's three dimensional pleasure-arousal-dominance (PAD) space in order to model the moods [8].

**Mood Update**

The three traits in Mehrabian's model are *pleasure* (P), *arousal* (A), and *dominance* (D) which are independent from each other, forming a three dimensional space. Pleasure-displeasure is related to the positivity or negativity of the emotional state, arousal-nonarousal shows the level of physical activity and mental alertness, and dominance-submissiveness relates to the feeling or lack of control. Eight different mood types are defined: *Exuberant, Dependant, Relaxed, Docile, Bored, Disdainful, Anxious and Hostile*. In the framework of INDIGO, the emotion engine receives the personality traits of the robot from the P-Server[1] and initializes the mood of the robot in terms of the OCEAN personality traits. The relationship between OCEAN and PAD is defined by Mehrabian [7]. According to our model, a specific personality results in a specific mood type and the mood range is restricted by the personality type. Mood update is also performed after the occurrence of an emotional appraisal. If there is an emotional state update with an emotion impulse from the DAM, first the emotional state is updated and then the mood is updated with the new emotional state. The conversion between OCC and PAD space is also defined by Mehrabian [8] and used and improved in [9].

**Emotional State Update:**

The OCC emotion appraisal model defines the emotions in relation to the events that cause them. For the INDIGO robot, the emotion engine receives an emotional appraisal of events from the emotional impact module and updates the emotional state and mood level accordingly. For more details of the mathematical model used by the emotion engine, we refer to [10].

**Emotion and Mood Decay**

Another important factor is the deterioration of the affective state as time passes. Emotions return to their normal state in a relatively short period after the occurrence of a triggering event. Decay is based on the intensity of the emotion and on personality. Personality defines an individual's control over his/her emotions. Decay is mainly related to the OCEAN trait "neuroticism", which represents emotional stability. For a neurotic person, positive emotions will disappear faster and negative emotions will disappear slowly, whereas the reverse holds for a more stable person. The decay of emotions and mood is modelled with an exponential curve.

---

[1] P-Server or Personalisation server has been developed by NCSR "Demokritos" and serves as the central storage server for all knowledge structures.

# 4 EXPRESSION OF EMOTIONS

Emotions are expressed in a multitude of ways, while being addressed at the visitor. Thus natural language generation and speech synthesis (spoken utterance), can be influenced to reflect the robots affective features. This can be realised by varying the voice tone or even by influencing the very formation of the robot's utterance. Moreover, emotional responses can be exhibited by the robot's head movement, by facial features and even by the robot's movement.

## 4.1. Speech Synthesis

Concerning the Indigo Robot's voice, the ACAPELA text to speech system (TTS) will be used. This is a multilingual high quality TTS recognized as one of the more natural on the market (http://www.acapela-group.com). It is a unit selection algorithm [13, 14] using a large speech corpus, which contains tens or hundreds of instances. Given a phoneme stream and a target prosody for an utterance, the TTS selects an optimum set of acoustic units which best match the target specifications. The result of such system is a highly natural synthetic voice that is very close to human voice.

Although such techniques produce speech that is very natural, it sounds as being very neutral, consequently no emotions can be conveyed using them. In order to add an emotional nuance to the ACAPELA TTS system, we have decided to record for two languages (English and Greek), a Unit selection corpus for each of the two emotions that will be used for the robot (Happy and Sad). Thus the chosen techniques are an off line recording of the standard corpus used for unit selection in happy and sad modes for the speaker. After those recordings, we will have for the same speaker 3 different voices (a neutral voice corresponding to the standard neutral voice that is being recorded, plus a sad and a happy voice). The system will be able to select one of the voices depending on the emotion that will be given by the Dialog Manager to the robot. It has also been decided to record a third voice for each speaker (for English and Greek) in order to enable the TTS system to put emphasis on specific part of the text. For example the robot would be able to express particularly important information to the visitor by putting emphasis on part of the sentences: "It is the best preserved temple of the Doric Style in Greece".

Thus for each specific emotion that the robot will express using the synthetic voice, the dialog manager will only have to select the adequate voice (neutral, happy or sad) and the TTS system will automatically pronounce the text with this specific emotion. In addition, it will be possible to put some emphasis on part of the sentences in order to put focus on important part of the sentences. The robot will thus be able to express specific emotions as a guide tour using a synthetic voice.

## 4.2. Facial Expression Synthesis

Facial expressions of the INDIGO robot head can be considered in terms of different components such as, lip-synchronized speech, face idle motions, emotional expressions, conversational expressions, and look-at behaviour.

Lip-synchronized speech is produced automatically by extracting phonemes from written text with the aid of ACAPELA TTS module. Emotional expressions are controlled by the emotion engine directly by converting OCC emotions to basic Ekman primitives (e.g. happy, sad, angry, fear, disgust and surprise). Conversational expressions such as, thinking, emphasizing, asking questions, and facial shrug, are constructed from tagged text originating from the DAM, and with timing information originating from the TTS system. Face idle motions such as, head movement and eye blinking are generated randomly to prevent repetition in animation. In addition, the robotic head is able to focus its face to a specific person or to a specific exhibit. This kind of behaviour is generated by moving the neck of the robot to the target position based on the data originating from the vision and the navigation modules.

It should be noted, that all of the aforementioned facial expression synthesis pieces of information are blended into a single stream in MPEG-4 Facial Animation Parameters (FAP), format and applied to the robot head. The MPEG-4 FAP to robot converter is being developed by MIRALab, at the University of Geneva.

# 5 CONCLUSIONS & FUTURE WORK

We have briefly described a robot platform with an affective component that interacts with the overall architecture in order to create and express emotional responses with a view to exhibiting intelligent behaviour.

This system will be eventually an integral part of a cultural institute for the benefit of the visitors; in particular visitors will interact with the INDIGO robot to learn about the exhibits. Moreover, improvements and extensions are planned in the individual modules of the INDIGO architecture. Further developments will also be effected following field studies, in which the INDIGO platform will be evaluated.

# ACKNOWLEDGEMENTS

# REFERENCES

[1] P.T. Cost and R.R. McCrae. Normal personality assessment in clinical practice: The NEO personality inventory. *Psychological Assessment,* 4(5-13) (1992).

[2] D. Galanis and I. Androutsopoulos. Generating multilingual descriptions from linguistically annotated OWL ontologies: the NaturalOWL system. In *Proceedings of the 11th European Workshop on Natural Language Generation (ENLG)*, Schloss Dagstuhl, Germany, pp. 143-146 (2007).

[3] A. Orthony, G.L. Clore, and A. Collins. The cognitive Structure of Emtotions. Cambridge University Press (1988).

[4] D. Traum and S. Larsson. The information state approach to dialogue management. In *Current and New Directions in Discourse and Dialogue*, eds. J.Juppenvelt and R. Smith. Kluwer Academic Publishers, the Netherlands (2003).

[5] S. Konstantopoulos, V. Karkaletsis, and C. Matheson. Robot personality: representation and externalization, In *Proc. Intl. Workshop on Computational Aspects of Affective and Emotional Interaction (CAFFEi)*, Patras, Greece, July (2008).

[6] D. Bilidas, M. Theologou, and V. Karkaletsis. Enriching OWL ontologies with linguistic and user-related annotations: the ELEON system. In: *Proc. 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, Patras, Greece, Oct. 2007, vol. 2, IEEE Computer Society (2007).

[7] A. Steed, E. Tanguy, X. Pan, C. Loscos, V. Vinayagamoorthy, M. Gillies and M. Slater. Building expression into virtual characters, *Eurographics State of the Art Reports* (2006).

[8] A. Mehrabian. Analysis of the big-five personality factors in terms of the PAD temperament model. *Australian Journal of Psychology*, 48(2):86--92, (1996).

[9] P. Gebhard. ALMA - a layered model of affect. In Proceedings of the *Fourth International Joint Conference on Autonomous Agents and Multi Agent Systems (AAMAS)*, pages 29-36 (2005).

[10] Z. Kasap, M. Ben Moussa, P. Chaudhuri and N. Magnenat-Thalmann. Making Them Remember - Emotional Virtual Characters with Memory. *IEEE Computer Graphics and Applications* (2008) [under review]

[11] S. Dupont and H. Bourlard. Multiband approach for speech recognition. *Proc. of ProRISC/IEEE Workshop on Circuits, Systems and Signal Processing*, pp. 113-118, Mierlo, The Netherlands (1996).

[12] S. Dupont, C. Ris. "Multiband with Contaminated Training Data", Proc. CRAC Workshop (EUROSPEECH satellite event), Aalborg, Danemark, Sept. 2001, (2001)

[13] A.J. Hunt and A.W. Black. Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 373-376. Atlanta, Georgia, (1996).

[14] T. Dutoit. Corpus-based Speech Synthesis. *Springer Handbook of Speech Processing*, Benesty, Jacob; Sondhi, M. Mohan; Huang, Yiteng (Arden) (Eds.), Springer, pp. 437-453 (2008).