Let's find a restaurant with Nestor
A 3D embodied conversational agent on the web !

Danielle Pelé, Gaspard Breton
France Telecom R&D
Human Interaction Division
Hyperlanguages and Multimedia Dialogues Lab.
35512 Cesson Sévigne Cédex, FRANCE
fax: 33 2 99 12 40 98, tel: 33 2 99 12 41 84
{danielle.pele, gaspard.breton}@francetelecom.com

Franck Panaget, Samuel Loyson
France Telecom R&D
Human Interaction Division
Dialogue and Intelligent Intermediation Lab.
Technolpole Anticipa, 22300 Lannion FRANCE
Fax: 33 2 96 48 30 50, tel: 33 2 96 05 28 52
{franck.panaget, samuel.loyson}@francetelecom.com

Abstract

This paper introduces a study on an interactive 3D Embodied Conversational Agent, Nestor, resulting from the integration of real time rational dialogue agent and 3D facial animation modules in a generic networked based architecture. It is adaptable on various networks and terminal platforms. In this paper focus is made on real time 3D facial animation and natural dialogue engines.

Keywords: 3D Embodied Agent, Real time Facial Animation, Natural dialogue engine

1 Introduction

Virtual creatures are now implemented on a lot of electronic trade websites. Market analysts foresee that the use of virtual humans as interfaces for websites and mobile portals will dramatically increase the rate of transactions and reduce costs.

Nevertheless the role of virtual characters is still quite passive: they provide a graphical representation in addition to text and audio but the relation with the user remains relatively weak. A lot of usage tests and in particular WoZ tests show that the users are expecting something more from these characters. They are demanding more interactivity, more transparency in communication. As reported by [Cas01a], a virtual character must be more than only "a pretty face !". People want to dialogue with virtual characters as if they were human beings with all the natural modalities we use to communicate, that is to say, mainly speech and gesture. The users want the characters to answer their questions in a very natural manner: with expressive speech and correlated non-verbal behaviors. In this paper we present such an Interactive 3D Embodied Conversational Agent.

The following section presents the overall architecture and sections 3 and 4 focus on the real time facial animation engine and dialogue manager. We will conclude with results and future works.

2 Overview of the system

Building a virtual human is a multi-disciplinary effort joining artificial intelligence, computer graphic technologies and a lot of knowledge from social science. As pointed out in [Gra02], researchers in each of these fields contributing to an ECA must understand the other disciplines and design their tools so that they can interoperate into a system that gives life to this interactive ECA.

To design a 3D ECA one can interact with in a natural manner, the required modules running in real time are at least:

- Speech recognition,
- Dialogue engine,
- Text To Speech (TTS),
- Animation engine.

We propose this kind of system where all the modules are embedded in a generic networked based architecture. Speech and dialogue processing are achieved on a server side while avatar animation engine and audio rendering are realized on the client side. The avatar animation is driven by speech phonemes provided by the TTS and behavior tags are provided by the dialogue engine. Tags and phonemes are part of a proprietary XML language

streamed to the client with the encoded synthetic audio. The player is embedded in an ActiveX, which can be easily integrated in a web page. The modules running on the client side have been designed in order to be able to run on PDAs or mobile phones. The stream conveyed to the client is very light and only requires small bandwidth.

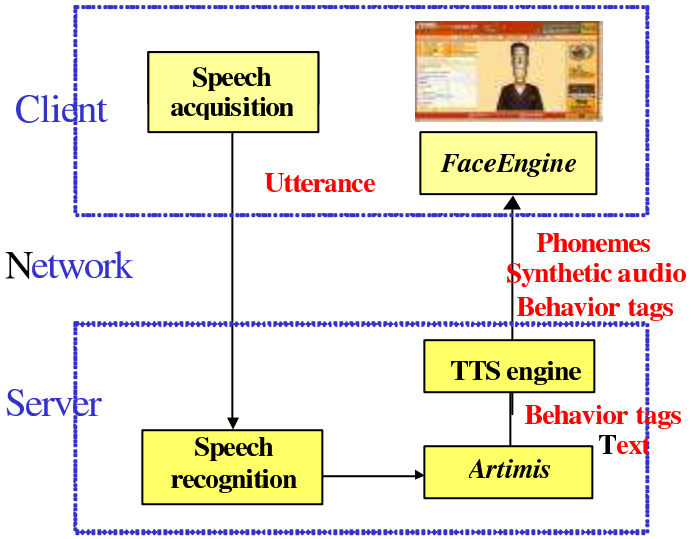

Figure 1: Overall system architecture.

For the moment, the server and client are running on Windows platforms. The next chapters emphasis will be on the real time 3D facial animation engine, *FaceEngine*, and the natural dialogue engine, *Artimis*.

3 Facial Animation

Facial Animation is made by an animation engine called *FaceEngine* [Bre00]. This animation engine is real time and has been designed in order to realize conversational agents. It is coupled with real time speech synthesis and voice segmentation. Scalability can also be achieved on the animation engine as well as on the meshes through the use of Dynamic Level Of Detail.

FaceEngine is a hybrid animation system using both muscular and parametric animation. Muscular animation is very interesting because it allows a very compact representation of the human expressions. It also provides a universal set of expressions making the creation of new faces straightforward.

The muscular system is based on the human anatomy and is made of 29 effector muscles . The set of effectors has been built in order to reproduce the main facial muscles. An effector muscle is a deformation module that moves the vertices in its area in influence in order to produce the effects of a muscle contraction.

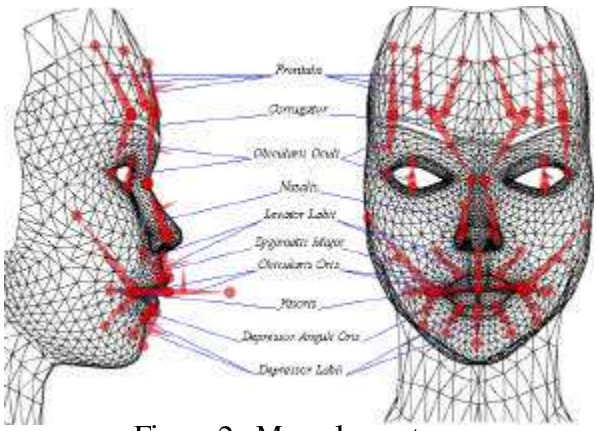

Figure 2 : Muscular system.

The effector muscles work somehow on the same principle as [Wat87]. Each effector has an area of influence and each vertex in this area is attracted toward the head of the effector when a contraction occurs. The effector contraction is computed in a way to simulate the skin elasticity.

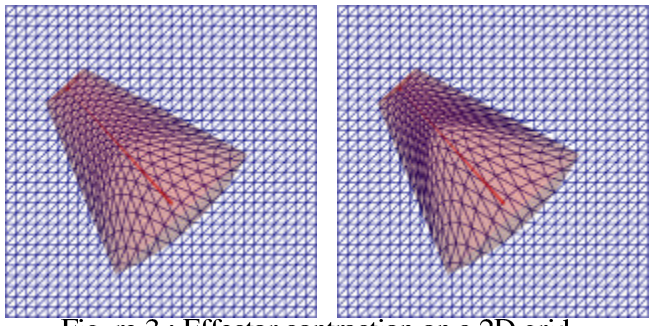

Figure 3 : Effector contraction on a 2D grid (at 40% and 60%).

Effectors are grouped together in order to compose the displacement vectors produced on the same vertex. This composition prevents the apparition of unnatural effects when vertices are moved too much. This composition is realized through the use of 8 composition modules.

All the movements that cannot be realized by the muscular system are achieved by a parametric system. This system is made of 12 modules :

- ?? 6 of them (jaw, neck, 2 eyes and 2 eyelids) compose the core of the parametric system ;
- ?? the 6 others (2 cheeks, 1 teeth lighting and 3 wrinkles) are considered as additional because they perform only aspect changes (bump mapping or texture blending).

So far, the animation engine is composed of 49 modules. At each frame, each of these modules takes a primitive called a *Control Unit* as input. These CU'scome from computations performed on the expression corpus or, at a lower level, can be sent directly to the animation engine.

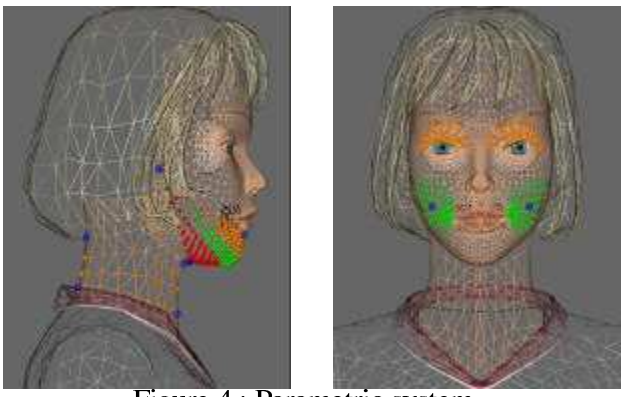

Figure 4 : Parametric system
(jaw, neck, eyelids, cheeks…).

FaceEngine also provides scalability so that the animation can best fit to the computing power of the target machine. The scalability engine works on the animation system itself by turning on and off some modules as well as on the 3D model through the use of Dynamic Level Of Details.

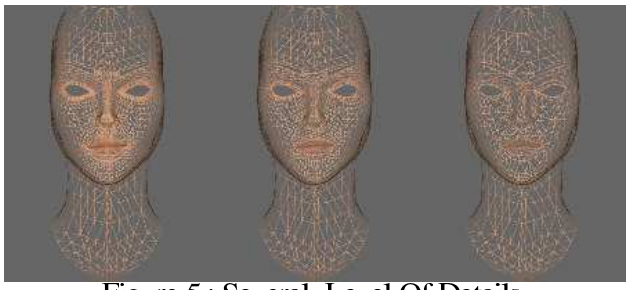

Figure 5 : Several Level Of Details.

Dynamic Level Of Detail is achieved in real time with vertex removal and/or edge collapse. Vertices are ordered according to the curvature and to their contribution to the animation system. The less valuables are the first to disappear.

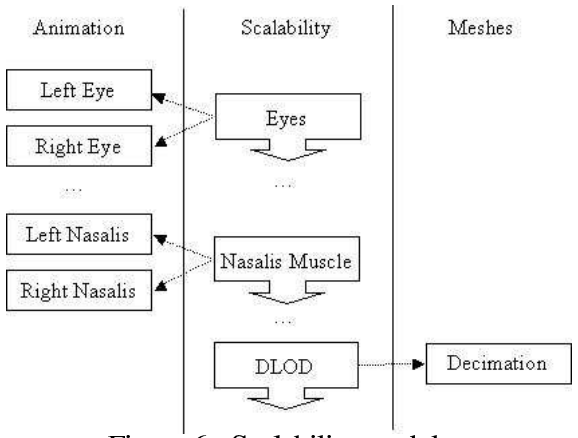

Figure 6 : Scalability modules.

The scalability system is organized around a sorted list of scalability modules to execute. Each module has in charge one or several animation modules or DLOD management. The adaptation process starts with the highest module. When the execution of the module found out it's not enough to gain power, the next module is executed and so on until there is no more modules. On the contrary, if there is remaining power, the system can go backward.

4 Natural dialogue engine

The natural dialogue engine *Artimis* provides a generic framework to instantiate intelligent dialogue agents [Sad97, Sad99]. These agents can interact with human users as well as with other software agents. In a context of human-agent dialogue, *Artimis* can engage with users in mixed-initiative cooperative interactions in natural language (contextual interpretation of user's inputs, negotiation ability, interaction flexibility, cooperative reactions, etc.).

Artimis is based on a theory of interaction: an integrated model of mental attitudes and rational actions. The theory, expressed in a logic of mental attitudes and actions, is formalized in a first-order modal language. The quantification domain contains actions, agents, and objects. The mental model of an agent is based on three attitudes: *belief, uncertainty*, and *choice*. The attitude of *intention* is defined as a composite concept defined in terms of belief and choice. In this paper, we use only belief (B_i? means agent *i* believes that ? is true") and intention (I_i? means agent *i* intends to bring about ? "). To recognize and plan actions, the logic involves, in particular, the operators *Feasible* and *Done*. The formula *Feasible(a)* means that action *a* can take place. The formula *Done(a)* means that action *a* has just taken place. In this framework, dialogue is modeled as the observation and the planning of communicative act(ion)s such as *request*, *inform*, *infirm* (e.g., <*i*,*inform*(*j*, ?)> is the act of agent *i* informing agent *j* that ? is true). Generic principles of rationality, communication and cooperation compose the theory of interaction.
A rational unit, the kernel of *Artimis*, implements this theory through an inference engine. This unit gives to *Artimis* its dialogue abilities, which results from an explicit reasoning process [Bre96].

When applied to human-agent natural dialogue, *Artimis* requires a natural language processing unit (interpretation and generation), and an interface management unit.
The user's utterance interpreter produces the best coherent interpretation (expressed in terms of a sequence of communicative acts) based on the word sequence output given by the speech recognizer. The natural language generator does the opposite by producing utterances that verbalize the plan of commu-

nicative acts planned by the rational unit [Pan97]. Interface management unit deals with multimodal fusion and fission. When *Artimis* is coupled with an avatar, the interface unit inserts additional information into *Artimis*' utterances in order to control avatar's behavior. The control takes place at two levels.

At a first level, it manages the global behavior of the avatar. For example, the avatar has a random head movement. When *Artimis* has no interlocutor; it looks ahead when it speaks to users or listens to them, and it looks down when it computes its reaction (i.e., the period between the end of user's utterance and the beginning of *Artimis*' answer).

At a second level, *Artimis* manages the local behavior of the avatar in order to be coherent with the content of its utterance. This control is based on *Artimis*' mental state and the sequence of communicative acts it has planned. A number of avatar behaviors can be associated with specific mental state patterns. Currently, we have identified close to fifteen patterns. Here are some examples of patterns (*s* denotes the system *Artimis* and *u* the user).

Pattern 1: yes-no question
When a/the user asks a yes-no question (1), the avatar shakes its head according to the fact that the answer is positive (2) or negative (3).

$B_s Done(<u,request(s, <s,inform(u,?)> | <s,inform(u,? ?)>)>)$ (1)

$B_s(I_s Done(<s,inform(u,?)>)$
$?Feasible(<s,inform(u,?)>)$ (2)

$B_s(I_s Done(<s,inform(u, ? ?)>)$
$?Feasible(<s,inform(u, ? ?)>)$ (3)

Pattern 2: satisfaction of user's intention
When *Artimis* believes that the user has an intention to know an object (1), then it is happy when the intention is satisfied (2) and unhappy otherwise (3).

$B_s I_u Bref_u(?x ?(x))$ (1)

$B_s(I_s Done(<s,inform(u, ?x ?(x) = object_A)>)$
$?Feasible(<s,inform(u, ?x ?(x) = object_A)>))$ (2)

$B_s(I_s Done(<s,infirm(u, ?y ?x ?(x) = y)>)$
$?Feasible(<s,infirm(u, ?y ?x ?(x) = y)>))$ (3)

Pattern 3: suggestion
When *Artimis* is not able to satisfy user's intention (1) but makes a suggestion (2a, 2b, 2c), the contrats is marked (between ? and ?) with forearm openings or head movements.

$B_s I_u ?$ (1)

$B_s(I_s Done(<s,infirm(u,?)>)$
$?Feasible(<s,infirm(u, ? >))$ (2a)

$B_s suggest(? , ?)$ (2b)

$B_s(I_s Done(<s,inform(u, ?)>)$
$?Feasible(<s,inform(u, ? >))$ (2c)

Moreover, *Artimis* marks the theme/rheme structure of (linguistic) clauses.
Finally, *Artimis* associates to each noun phrase, pronoun or proper name, the object it was intended to refer to and specifies if the object is introduced for the first time in the dialogue or not (*new* versus *evoked*). This information, associated with application specific avatar behaviors, can be used in the decision to produce physical action, such as arm's gestures and pointing.

All these pieces of information are represented in terms of XML tags in *Artimis*' utterances. Here is an example of dialogue with extended outputs.

U: I'd like a Greek restaurant in the twentieth district.
S: <UTTERANCE SPEAKER="Nestor", HEARER="user" SCENE="plan-resto"> <CLAUSE> You want a Greek restaurant in <OBJECT ID="dist20" CLASS="District"> the 20th arrondissement </OBJECT>. </CLAUSE> I've found 4. To continue, specify the location, consult the solutions or proceed to a new request.
U: I'd like to consult the solutions
S: <UTTERANCE SPEAKER="Nestor", HEARER="user" SCENE="plan-resto"> <CLAUSE> <THEME> The first restaurant is </THEME> <RHEME> <OBJECT ID="resto1026" CLASS="restaurant"> <NEW> L'Amarante </NEW> </OBJECT>. </CLAUSE> Do you want to have more information, hear about the next restaurants or make a new request?
U: and a Chinese.
S: <UTTERANCE SPEAKER="Nestor", HEARER="user" SCENE="plan-resto"> <CONTRAST ID="0"> <CLAUSE> There isn't any Chinese restaurant in the 20th district. </CLAUSE> </CONTRAST> However, <CONTRAST ID="0"> there are… </CONTRAST>

5 Results

A prototype has been realized for searching for restaurants in Paris. The user asks through a microphone for a restaurant and Nestor proposes the possible items while related web content is dynamically displayed according to the answer (map, pictures, menus…).

Users also have access to a written dialogue history, and can use text or mouse input, in addition or as an

alternative to voice. This can help people with impaired hearing or elocution, or be useful in noisy environments.

Figure 7 : Screenshot of Nestor.

6 Conclusion and perspectives

In this paper we presented a prototype of a 3D Embodied Conversational Agent with a focus on facial animation and natural dialogue engines. As stated in the introduction, people want these agents to behave as humans. Nestor is a first step and some tests show that people like it! Of course much work remains to be done to make the human-ECA relation more natural and characters more believable. Progresses have to be achieved in each technology contributing to the 3D agent system. It is also needed to enhance the approach by interpreting other user's modalities jointly to speech (gesture, emotion in face and speech…).

Concretely, our on going work concerning Animation and Dialogue now focuses on:

?? Adding new non verbal behaviors with hand gestures related to the dialogue context,

?? Increasing the ability of *Artimis* to control the avatar's behavior. First experiments have been done on associating *Beat* [Cas01b] and *Artimis* to provide richer behaviors markers.

7 Acknowledgement

The authors thank Justine Cassell from MIT/ Medialab for her helpful advice and discussions. They also thank their colleagues involved in the project: Paul Bagshaw and Thierry Moudenc (TTS), Denis Jouvet and Geraldine Damnati (Speech recognition), Benoit Simon and Romain Le Crom (Integration).

References

[Bre00] G. Breton, C. Bouville, and D. Pelé, "FaceEngine : A 3D Facial Animation Engine for Real Time Applications," presented at Web3D Symposium, Paderborn, Germany, 2000.

[Bre96] Bretier P. & Sadek D., A rational agent as the kernel of a cooperative spoken dialogue system: implementation of a logical theory of interaction, *Intelligent Agents III*, J.P. Müller, M.J. Wooldridge, N.R. Jennings eds., LNAI, 1996.

[Cas01a] Cassell, J., Bickmore, T., Vilhjalmsson, H., Yan, H. (2001). "More Than Just a Pretty Face: Conversational Protocols and the Affordances of Embodiment." *Knowledge-Based Systems* 14: 55-64.

[Cas01b] Cassell, J., Vilhjálmsson, H., Bickmore, T.(2001) "BEAT: the Behavior Expression Animation Toolkit." *Proceedings of SIGGRAPH '01*, pp. 477-486. August 12-17, Los Angeles, CA

[Gra02]]Jonathan Gratch, Jeff Rickel, Elisabeth Andre, Norman Badler, Justine Cassell, Eric Petajan (2002) "Creating Interactive Virtual Humans: Some Assembly Required" *IEEE Intelligent Systems* 17(4): 54-63

[Pan97] Panaget F., Micro-planning: a unified representation of lexical and grammatical resources, *6th European Workshop on Natural Language Generation*, Germany, 1997.

[Sad97] D. Sadek, P. Bretier et F. Panaget, ARTIMIS: Natural dialogue meets rational agency, *Proceedings of the 15th International Joint Conference on Artificial Intelligence* (IJCAI'97), Nagoya, Japon, pp. 1030-1035, 1997.

[Sad99] Sadek D., Design considerations on dialogue systems: from theory to technology – The case of Artimis, *ESCA TR Workshop on Interactive Dialogue for Multimodal Systems*, Germany, 1999.

[Wat87] K. Waters, "A Muscle Model for Animating Three-Dimensional Facial Expression," presented at Proceedings Of Siggraph, Anaheim, California, 1987.