VIRSTORY: Using Human Behavior Studies to Design a Multimodal Collaborative Virtual Environment

Atman Kendira[1,2], Laurence Perron[1], Sébastien Carbini[1],
Jean-Emmanuel Viallet[1], Lionel Delphin-Poulat[1]

[1]*France Telecom R&D Lannion*
2 avenue Pierre Marzin
22307 Lannion – France

[2]*HEUDIASYC Laboratory, CNRS / UTC*
University of Technology of Compiègne
60205 Compiègne – France

{atman.kendira, laurence.perron, sebastien.carbini,
jeanemmanuel.viallet, lionel.delphinpoulat}@francetelecom.com

Abstract

In a Collaborative Virtual Environment, every distant user is able to meet other users to work, to communicate and to play together. France Telecom R&D have developed a multimodal Collaborative Virtual Environment named "VIRSTORY", a digital storytelling environment using bimodal interaction (speech and gesture) and where every user is represented by an autonomous 3D character specially used to support user's non-verbal behavior like gestures or facial expressions.

This paper describes the design process of this multimodal Collaborative Virtual Environment from the results of Human behavior studies.

1. Introduction

One of the most important roles played by technologies are connecting people and mediating their communications. Building technology that mediates communication presents a number of challenging research and design questions. Apart from the fundamental issue of what exactly gets mediated, two of the more crucial questions are how the person being mediated interacts with the mediating layer and how the receiving person experiences the mediation.

This paper is concerned with both of these questions and proposes a theoretical framework of mediated conversation by means of automated characters.

The non-verbal part of the communication is very important in a face-to-face conversation. It should increase the quality of the collaborative task if it can be provided an animated feedback of the non-verbal aspects of human interaction and communication. A video-based solution offers a part of the non-verbal aspects, however, a video is not easy to manipulate and does not provide the same synthesis capacities as a synthetic character.

In this paper first, we investigate a way to permit the "natural" mediated communication between distant users in virtual environment and especially with creative tasks where the non constraint interactions are essential. Next, we present the VIRSTORY platform architecture and in section three we describe the speech and gesture multimodal man-machine interface, allowing users to interact remotely. Finally, we present the avatar architecture and in particularly we explain the intelligent part of our system able to automate animation avatar according to distant users' actions or behaviors in VIRSTORY.

2. Humans Behavior Studies

Our aim is to carry out some experiences in order to design a multimodal Collaborative Virtual Environment (CVE) and particularly storytelling environments. Our experimentations concern Human/Human mediated collaboration without intrusive devices. Our recent studies are focused on virtual collaborative games with small groups of players. In these particular games, the users are represented by characters or avatars in CVE and it is primordial to animate the characters in a way that they are supported by the human communication especially non-verbal behavior: gestures, gaze, life sign, social attitudes, emotion...

Voice and gesture are both the basis of natural dialogue between players and the bimodal interaction support of the digital storytelling setup. This double function is very interesting for the users but remains a Human Computer Interaction (HCI) research field and

more specifically "new multimodal humanized interfaces" [1-3].

In fact, we are focusing on voice and gesture interaction because these modalities give the user more freedom to create and participate in the story narration [4]. We are specifically experimenting Human behavior in the storytelling task.

2.1. "Once upon the time..."

We have imagined a creative digital game that anybody without particular competences, can play, a narrative game called "*Once upon a time*". This game uses pieces of the story (cards) e.g.: princess, wolf, etc. Each user must play with her/his pieces of story and with the other users to elaborate a marvelous and coherent story. This game was the starting point of the VIRSTORY CVE design to support storytelling task. If the primary aim is to conceive VIRSTORY CVE, the secondary aim is to understand Human' behavior especially the non-verbal behavior like gaze, facial expression, and gesture of hand to animate a character in a future creative CVE [5].

Existing character animation techniques in CVE prompted us to study sensorial deprivation along two lines of reflection. We wished to establish how subjects interact in a trio, and what they do when they are deprived of a form of expression (eyes, hand, and face) as the case with the intrusive device in CVE (Figure 1). The idea was that subjects compensate for the absence of one of these forms of expression either verbally or by means of other non-verbal behaviors not under constraint in the experimental situation.

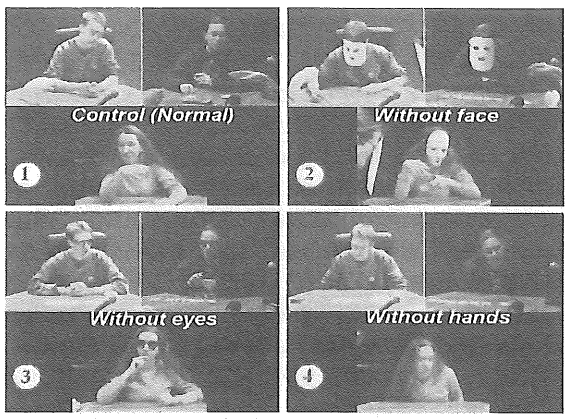

Figure 1. Players during storytelling experiment

Observations of this experiment showed that character can already be animated with recurring self-contact gestures. These gestures are resulting from behaviors learned for "adaptation" purposes, such as washing oneself, but used in an interaction without these 'adaptive' needs being present. Moreover, these gestures contribute to the feeling of presence that users expect when facing a character without speech analysis. Unfortunately, self contact does not really seem to support efficient communication between users. According to [6], this non-verbal behavior concerns more the internal sensation than the communicative intention.

Self contact gestures are used to create an animation library that can be played by characters during a similar situation. One must nevertheless bear in mind the fact that mechanisms of mutual adjustment exist, from the point of view of non-verbal behavior that cannot reduce a human's behavior to that of an animated character.

2.2. Speech and voice experimentation

We focused on users engaged in a collaborative storytelling task in which gestures and speech were essential as a medium for the collective narration: the human communication and objects manipulation (e.g. cube). We wished to establish how subjects interact in a duo, e.g.: face to face, and what they do when they use hand gesture or/and speech to communicate with the system or/and with the distant partner. The aim is to improve multimodal interfaces supporting mediated collaboration and particularly natural human communication between human and human. (e.g. non-verbal behavior which implies non intrusive devices).

In VIRSTORY game (Figure 2), the user is represented by a virtual character. When the user takes a cube, the distant user can see the character with the cube and the colored feedback on the cube. If the character is red then the cube is red, etc.

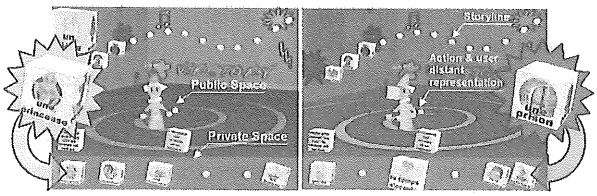

Figure 2. VIRSTORY: on the left, the first player-on the right, the second player screen

Several subjects (between 10 and 30 years old) participated in an experiment in which the instruction was to "create the most attractive and coherent story". We carried out a "Wizard of Oz" experimentation; an experimenter perform the user action and the user played with an associate partner. The role of the associate partner is important because he plays with the subject and sometimes, he encourages the subject to use his vocal incentives.

Figure 3. Left: a user with a laser pointer. Right: the WoZ interpreter carrying the user's command and the partner in the background playing with the user

This experimentation allows understanding the organization of the oral and gesture modalities between themselves, as a function of the habits and competences of the subject (adult/child difference for example). One of the interests of multimodal interfaces is to let a user choose the way to interact as she/he prefers. If pointing with a laser is efficient when this tool is given to a user, the question remains open whether such efficiency is kept without a physical pointing device, for example with computer vision pointing gesture recognition. The interpreter only reproduced gestures performed on objects of the storytelling setup, while other gestures such as wrist or arm rotation to turn a cube were interpreted.

Concerning speech-gesture interactions, the verbalizations of the subjects and the images from recorded video indicate addressability problems. In the experiments, subjects speak aloud to themselves; introducing meta-discursive items such as "princess, I had a princess, a queen, **I should have put the princess**". The subjects speak to the machine ("I will talk to the machine") and say so the system and/or the partner ("wood", "jewels"). The interpreter uses continuously the discursive context to understand who the player is talking to.

To conclude, such an experiment allows defining a spoken vocabulary, to observe, depending on the context, the importance of a given modality over other modalities. The most important lesson learnt from this experiment is that modality influences the type of cooperation. The more the subjects pointed and the lesser the story was built in a cooperative manner: the story was build by successive elements without real integration of the distant user.

Cooperation between the two users increased as users performed less manipulation and remained concentrated on the narrative communication (this was specifically observed for the users who relied on speech-gesture interactions). It can be observed that for the voice only subject, the number of oral commands was low in order to favor the collaborative narration with the distant user. Only one female subject had to rely more than the other subjects on her memory (to remember the content of the cubes) owing to the small number of commands used.

Multimodality involves more than signal (vision and speech) processing, synchronization, fusion and session techniques, user and context awareness and HCI. Indeed, it was noticed, during the WoZ experiments, that the experimenter not only conducted modality fusion, but had to deal with discursive and game context integration, to interpret the ambiguities of the request, anticipate, cope with difficulties encountered by users and induced by the interface. The human interpreter response continuously adapted to the user which is yet far away from any available system behavior.

Finally, the lessons learned from the sensorial deprivation and WoZ VIRSTORY experiments are that the technical problems encountered during the conception of a multimodal collaborative interface should not shadow other questions: for which task, for which user, for which type of interaction or cooperation is the system designed. The impact of multimodality should be further investigated when moving from human-computer interaction towards Human-Human (with machine mediation) collaboration.

Moreover, the analysis that we carry out here points out the necessity to work on human studies particularly in situ or in ecological conditions, in order to discriminate transpositions from the real world into the virtual world. When transposing an activity and its objects into a virtual world, we create abstractions, new functions, new tools and new habits. We then have to think, model and design the tools and the functions taking into account all aspects: the way they are used, the context, the habits, the usage, and the cognitive skill, dexterity or social aptitude of the different users.

3. The VIRSTORY 3D CVE

VIRSTORY platform integrates user's non-verbal and communicative behaviors. This platform is based on client/server architecture. Each client is responsible for rendering a single user's view. All users connected to the same embedded server see each other's characters. A character is a 3D full body. The user is able to:

- Play in 3D synchronous distributed platform with other players.
- Speech with other users to design a story.
- Manipulate (with speech-gesture modality) 3D cubes without intrusive devices.
- Observe users' non-verbal behavior via characters.

Figure 4 illustrates the architecture of VIRSTORY. The gesture and speech input modalities are interpreted by MOWGLI module (describe in next part) [4] and used as synchronous or asynchronous interface commands in order to

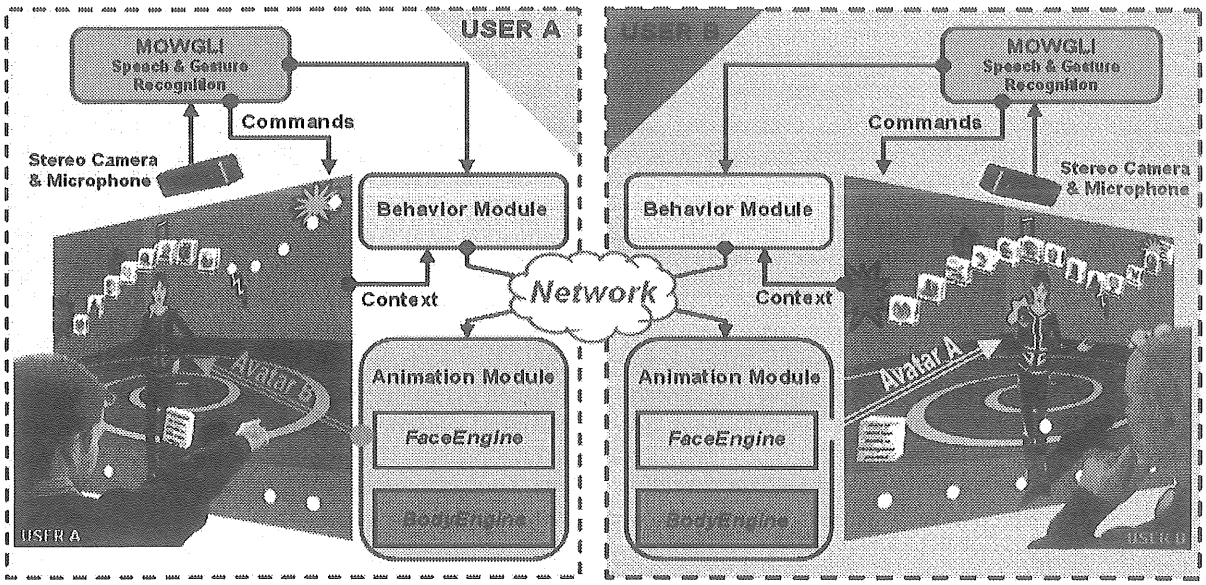

Figure 4. The VIRSTORY platform: User A see user's B character and vice versa

interpret and sometimes disambiguate the user's commands.

Behavior Module receives interpreted data (head and arms position, speech detection, speaking time, speech recognition) from MOWGLI Module and selects appropriate behaviors reproduced on distant user's character. Once a behavior or action is selected, the Behavior Module sends through network a behavior message (with character identity, animations to play and parameters animations) to others users' Animation Module of collaborative session.

Locally, all Animations Modules receive the behavior message from distant Behavior Module and activate the corresponding face and body animations on the character.

The user's voice is transmitted directly via network to another user and spatialized in our application.

4. The MOWGLI Module

MOWGLI (Multimodal Oral With Gesture Large display Interface) fuse and interpret speech and gesture input modalities taking into account the application context, allowing the user to interact efficiently with large display while moving freely in the room.

4.1. Gesture recognition

Most people instinctively use the eye-tip of the finger line to point at a target [7]. We use this convention in the present framework to estimate the pointing direction of the user.

Head and hands are detected and tracked [8] (Figure 5) utilizing the 3D data provided by a stereo camera. The face of the user is automatically detected by a neural network more precisely described in [9]. We detect the hands as a skin colored moving area in front of a vertical plane passing in front of the head (i.e. the hands have to be in front of this plane so as to be recognized, since everything behind the head is discarded) (Figure 5 left).

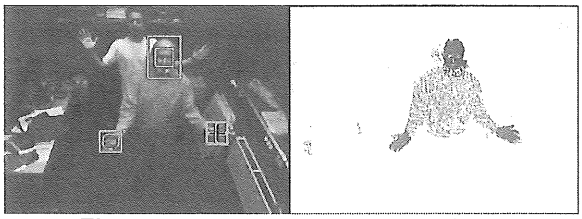

Figure 5. Head and hands tracking

The first detected hand is tagged as "pointing hand" and the second as "control hand". The system can be used by right-handed persons as well as by left-handed persons (predominant hand is generally used to point) without differentiating explicitly the right hand from the left.

Once detected, we track body parts (head and hands) simultaneously. The tracking process aims at representing each new image by a statistical model utilizing the EM algorithm [8]. The statistical model is composed of a color histogram and a 3D spatial Gaussian function for each tracked body part (Figure 5 right).

During runtime, pointing recognition is permanent. We use the axis, obtained from the first hand, and the 3D position of the head to compute the pointing direction. When the user utters "selection" the pointed 3D cube in VIRSTORY is selected. Once selected, using the "move" command, the 3D cube keeps moving, following the hand's motion to displace it in corresponding area. Finally, all actions

stopped when "O.K." is uttered. The user can also move a cube, by pointing the destination location and uttering the cube label (visible or hidden).

4.2. Speech recognition

To recognize speech commands, the speech signal is linearly sampled at 8 kHz in 16 bits. Next, we compute MFCC (Mel Frequency Cepstrum Coefficients) coefficients, each 16 ms, on 32 ms signal frames. The recognition system uses the frame energy, 8 cepstral coefficients and an estimation of the first and second order derivatives of the speech signal. Thus, the observation vector has 27 dimensions.

In the decoding system we use Hidden Markov Models. The recognized sentences syntax is described in a grammar. The used vocabulary consists of 100 words. Dependent on the context, each word is obtained by phonetic units concatenation named allophones [10]. The system finally outputs the n-best results [11].

We filter the input signal using a noise/speech detector component to provide the decoder only with speech signal surrounded by silent frames. Beginning of detection is not causal. The detection component provides several frames which precede the speech detection decision. But as the decoder is faster than real time, it recovers from the non-causality of the detection component.

To detect end of speech, some consecutive silent frames must be observed. Thus, the best solution can be provided as soon as the last frame has been received. Computing the n-best solutions generate a negligible lag compared to the lag due to the silent frames. The number of frames to detect the end of speech is a parameter of the noise/speech component and is set to 15. The lag between the end of speech and the result of the recognition is thus 240 ms, since we grab frames each 16ms. These times include the start and end silent frames which differ from the times of start and end of speech. The former can be computed from the noise speech parameters.

5. The Character Animation Module

VIRSTORY Animation Module is similar to that proposed by Blumberg [12] in that it consists of articulated character that is manipulated in real time, an Animation Module which utilizes descriptions of macro animated actions (such as self-contact gestures, beat, or some facial expressions) to manipulate the articulated character, and a Behavior Module which is responsible for higher-level capabilities, (such as generate various gestures, or engaging another actor in a conversation), and decisions about which animations to trigger. In addition, the Behavior Engine maintains the internal model of the actor, representing various aspects of an actor's moods, goals and personality.

5.1. Facial animation

To animate character's face, we use a face animation engine called "FaceEngine" [13]. This animation engine is real time and has been designed in order to realize conversational agents. FaceEngine is a hybrid animation system using both muscular and parametric animation. It is coupled with the Behavior Module which analyses user behavior (speech detection, words recognition and user's head position) in order to generate the appropriate face animation for the distant character. The main functions include eyes, eyelids and neck movements, facial expressions (anger, disgust, fear, happiness, sadness, surprise, etc), behaviors and automatic life signs.

During a session when a user speaks on his/her microphone, we animate the lips of his/her distant character. The main difficulties are the synchronization of lips with the speech and the necessity to make full phoneme segmentation in real-time.

When speech is detected, we also play automatic life signs (e.g.: eyes blink) and random visemes.

When several users are in the environment, this simple lips animation provides sufficient information to know who is talking. Even if facial animation is not precise enough to read the lips, it still provides a strong, essential non-verbal communication feedback.

5.2. Body animation

Observations in real situation of collaboration reveal that some signification gestures or actions are recurring [14]. It is interesting to create a library of relevant pre-computed poses and animations that can be replayed automatically, randomly or voluntarily during the collaborative session.

To realize this library of body animation, we have developed an authoring-tool named "BodyEngine" able to generate animations stocked in a data base of the character according to Behavior Module which selects the appropriate actions.

An animation is defined as a sequence of poses associated with a time. Our animation engine is able to go from one pose to another with several means of interpolations and support some functions such as animations blending or time deformation.

6. The User Behavior Module

The Behavior Module constitutes the mind of the characters. It drives automatically the distant

character animation according to rules defined by the application builder.

We propose a generic and qualitative approach based on a multi-agent system (one agent for one user) allowing the definition of procedures dedicated to behaviors characters. The resulting agents are reactives but the behavior's specification is cognitive.

Reactive approaches which link the sensors values to actuators values by functions without memory [16]. The reaction of an agent to a modification of its environment is then fast (e.g.: labialization of the character).

Cognitive approaches come from classical artificial intelligent [15]. It is based on the assumption that the agent's perception and reasoning must be described by symbols and rules. Such approaches are highly declaratives and allow the definition of inferences rules which manage the execution of agent's actions.

7. Conclusion & Future work

Our objective is to preserve the advantages of "face to face" communication in the mediated collaboration and increase the potentialities of multimodal interfaces which are supporting the natural human communication (e.g. non-verbal behavior that implies the non intrusive devices like "hand free pointing").

As it has been previously demonstrated non-verbal communication is important in face-to-face conversation. According to this observation we have developed a multimodal CVE named VIRSTORY, able to increase effectiveness of the virtual collaboration.

Figure 6 presents 2 snapshots of first results of the VIRSTORY platform with two autonomous characters.

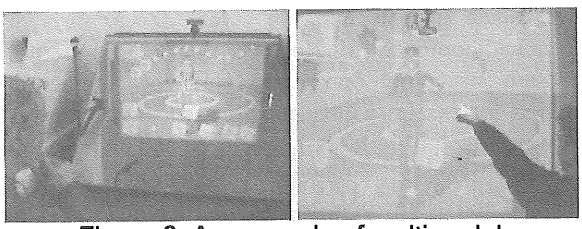

Figure 6. An example of multimodal collaborative storytelling game in VIRSTORY between two distant users

Based on first promising results we will focus our future work on one main point. We would like to improve our Behavior Module to take into account more non-verbal gestures like beats, self contact gestures, facial expressions, etc.

Future experiments with users should prove the efficiency of our choices.

8. References

[1] S. Buisine, J.C. Martin and N.O. Bernsen, "*Children's gesture and speech in Conversation with 3D Characters*", International Conference of Human Computer Interaction, Las Vegas (USA), 2005.

[2] S. Oviatt, P. Cohen and W. Lizhong, "*Designing the use rinterface for multimodal speech and gesture application: State of the art systems and research direction*", Human Computer Interaction, Addison Wesley Press.

[3] R. Prada and A. Paiva, "*Believable groups of synthetic characters*", International Conference on Autonomous Agents, Utrecht (Netherlands), 2005, pp. 37-43.

[4] S. Carbini, L. Delphin-Poulat and L. Perron, "*From a wizard of Oz experiment to a real time speech and gesture multimodal interface*", Signal Processing, 2005.

[5] L. Perron, "*An avatar with your own gestures*", Interact 2005, Rome (Italy), 2005, pp. 12-16.

[6] P. Ekman and W.V. Friesen, "*The Repertoire of Nonverbal Behavior: Categories, Origins, Usage, and Coding*", Semiotica, 1969.

[7] K. Nickel, E. Seemann and R. Stiefelhagen, "*3d-tracking of head and hands for pointing gesture recognition in a human robot interaction scenario*", IEEE International Conference on Automatic Face and Gesture Recognition, Seoul (Korea), pp. 565 -570.

[8] S. Carbini, J. E. Viallet and O. Bernier, "*Suivi statistique simultane des parties du corps pour des interactions bi-manuelles*", ORASIS, Fournol (France)

[9] R. Feraud, O. Bernier, J. E. Viallet and M. Collobert, "*A fast and accurate face detector based on neural networks*", Pattern Analysis and Machine Intelligence, vol. 23, no. 1, 2001, pp. 42-53.

[10] K. Bartkova and D. Jouvet, "*Modelization of allophones in a speech recognition system*", ICPhS (Internatioanal Congress of Phonetic Science), Aix-en Provence (France), pp. 474-477.

[11] R. Schwartz and Y. L. Chow, "*The n-best algorithm: an efficient and exact procedure for finding the n most likely sentence hypothesis*", ICASSP (International Conference on Acoustic Speech and Signal Processing), Albuquerque (USA), pp. 81-84.

[12] B. Blumberg and T. Galyean, "*Multi–Level Direction of Autonomous Creatures for Real–Time Virtual Environments*", Computer Graphics, SIGGRAPH '95 Proceedings, vol. 30, no. 3, 1995, pp. 47-54.

[13] G. Breton, C. Bouville, and D. Pelé, "*FaceEngine a 3d Facial Animation Engine for Real Time Applications*", In Web3D, 2001, pp. 5-22.

[14] L. Perron, "*What kind of gesture to animate an avatar?*" Lannion, France, France Telecom R&D, Internal Report, 2004.

[15] M. Igrand, M. georgeff, and A. Rao, "*An Architecture for real-time reasoning and system control*", IEEE Expert,1992, pp. 34-44.

[16] P.Maes, "*The dynamics of action selection*", International Joint Conferences on Artificial Intelligence, 1989, vol. 2, pages 991-998.