Communicating with virtual characters

Nadia Magnenat Thalmann, Prem Kalra, Marc Escher
MIRALab, CUI
University of Geneva
24, rue du General-Dufour
1211 Geneva, Switzerland
Email: {thalmann, kalra, escher}@cui.unige.ch
URL: http://miralabwww.unige.ch/

ABSTRACT

This paper sketches an overview of the problems related to the analysis and synthesis of face to virtual face communication in a virtual world. We describe different components of our system for real-time interaction and communication between a cloned face representing a real person and an autonomous virtual face. It provides an insight into the various problems and gives particular solutions adopted in reconstructing a virtual clone capable of reproducing the shape and movements of the real person's face. It includes the analysis of the facial expression and speech of the cloned face, which can be used to elicit a response from the autonomous virtual human with both verbal and non-verbal facial movements synchronised with the audio voice.

Keywords: **Virtual Human, Virtual Face, Clone, Facial Deformation, Real Time Facial Animation, Autonomous Virtual Human, Virtual Dialog, Phoneme Extraction, 3D Feature Points.**

1. Introduction

This paper is an account of a face to virtual face interaction system where a clone, representing a real person, can dialog with another virtual human, who is autonomous, in a virtual world. The dialog consists of both verbal and other expressive aspects of facial communication between the two participants. Section 2 gives an overview of the problem and describes major contributions related to the different aspects. Section 3 concentrates on our system and describes different components of the system. Section 4 presents issues related to the standardization of parameters for defining the shape and animation of the face. Future trends are outlined in the concluding remarks.

To clone is to copy. In our context, cloning means reproducing a virtual model of a real person in the virtual world. Here, our interest is restricted to one component of human figure, the face. The face is the most communicative part of a human figure. Even a passive face conveys a large amount of information, and when it comes to life and begins to move, the range of motions it offers is remarkable: we observe the lips, teeth, and tongue for speech, eyes and head movements for additional elements of dialog, and flexing muscles and wrinkle lines for emotions.

Developing a facial clone model requires a framework for describing geometric shapes and animation capabilities. Attributes such as surface color and textures must also be taken into account. Static models are inadequate for our purposes; the model must allow for animation. The way facial geometry is modeled is motivated largely by its animation potential.

Even though most faces have similar structure and the same set of features, there is considerable variation from one individual face to another. These subtle and small differences make the individual face recognizable. In modeling the face of a real person, these aspects have to be captured for the model to be identifiable as a clone.

Prerequisites for cloning a face are analyses of several aspects necessary for its reconstruction: its shape, and its movements due to both emotions and speech. This requires techniques from various fields. Shape includes geometrical form as well as other visual characteristics such as color and texture. Input for shape reconstruction may be drawn from photographs and/or scanned data. The synthesis of facial motion involves deforming its geometry over

time according to physical or ad hoc rules for generating movements conveying facial expressions and speech. The input for the facial motion of the clone will be the facial expressions and/or the speech of the real person.

2. Facial communication

Facial communication among virtual humans has recently attracted much attention. Cassell et al.[1] describe a system which generates automatic speech and gestures, including facial expressions, for modeling conversation among multiple human-like agents. This system, however, does not include cloning aspects. Thorisson[2] presents a mechanism for action control, using a layered structure, for communicative humanoids. The virtual environment consists of Gandalf, a simplified caricatural face, used as the autonomous actor.

There can be four different situations for face to virtual face communication in a virtual world. These are: real face to virtual face, cloned face to cloned face, virtual face to virtual face, and cloned face to virtual face. The four cases are briefly described as follows.

2.1 Real Face to Virtual Face

Here a real face communicates with a virtual face who is autonomous and has some intelligence to understand what the real face conveys. This may have two types of input from the real face: video input from the camera and speech from audio (microphone). The emotions of the real person are recognized by facial features extraction and tracking from the video input and the audio is converted to the text and phonemes. These are used by the autonomous virtual face to respond in a believable manner through his/her facial expressions and speech (see Figure 1). Mimicking the input from the real face video or speech are the special cases of this situation.

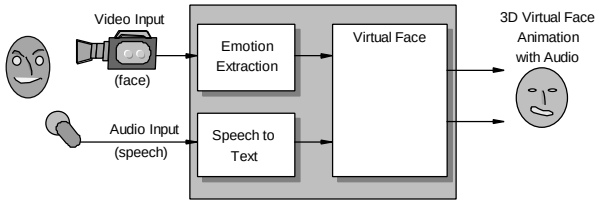

Figure 1: Real face to virtual face

2.2 Cloned Face to Cloned Face

In a networked 3D virtual environment communication may exist between the participants located at different sites represented by their 3D clones. This requires construction of 3D clone and facial motion capture of the real face from the video camera input for each participant. The extracted motion parameters are mapped to the corresponding facial animation parameters for animating the cloned models. The audio or speech input is also reproduced on each clone. Figure 2 shows the modules for such a communication.

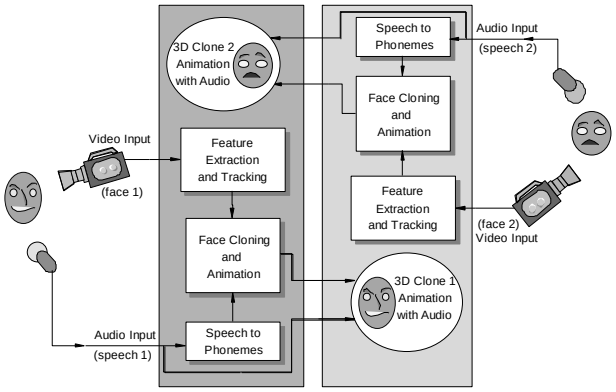

Figure 2: Cloned face to cloned face.

2.3 Virtual Face to Virtual Face

A virtual face who is autonomous may communicate with another autonomous virtual face. The communication may involve both speech and facial emotions (Figure 3). This situation does not require any cloning aspects. The autonomy of the virtual humans gives them the intelligence and capacity to understand what is conveyed by the other virtual human and generate a credible response.

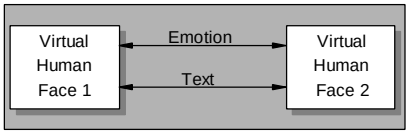

Figure 3: Virtual face to virtual face

2.4 Cloned Face to Virtual Face

A virtual environment inhabited by the clones representing real people and virtual autonomous human, would require communication between a cloned face and virtual face. This needs the cloning and mimicking aspects to reconstruct the 3D model and movements of the real face. The autonomous

virtual face is able to respond and interact through facial expressions and speech. This is of particular interest when there is more than one real participant being represented by their clone with one or more autonomous virtual humans. Although the participant may not be interested in watching their clone, they would be able to see the presence of the other real participants. Figure 4 shows the configuration of the communication between a cloned face and a virtual face.

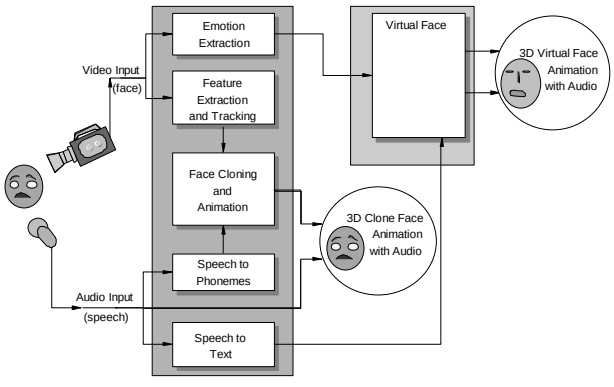

Figure 4: Cloned face (of real face) to virtual face.

The communication between a cloned face and virtual face addresses all the problems involved for facial cloning and communication and is considered as a general case for all the above four situations.

We now describe our system, with its different modules, that enables face to virtual face communication in a virtual environment inhabited by the virtual clone of a real person and the autonomous virtual human. The system concentrates only on the face. Other body parts, though often communicative in real life, are for the moment considered passive. Although the system is described considering one cloned face and one virtual autonomous face, it can have multi cloned and multi virtual faces.

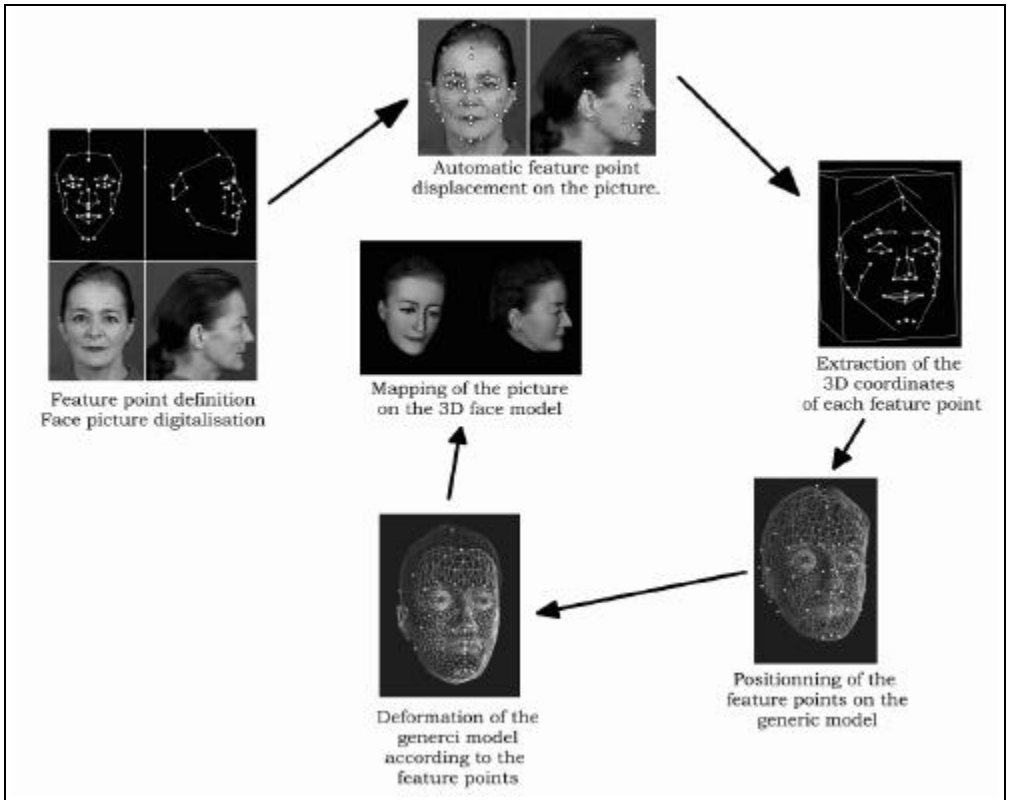

Figure 5: Steps for clone construction.

3. A description of our system

The general purpose of the system is to be able to generate a clone of a real person and make it talk and behave like the real person in a virtual world. The virtual world may contain another virtual human, who is autonomous; dialog and other communication can be established between the clone and the autonomous virtual human. The autonomous virtual human understands the expressions and speech of the real person (conveyed via the clone) and communicates using both verbal and non-verbal signals. The dialog is simulated in a 3D virtual scene

which can be viewed from different remote sites over the network.

Implementing such a system necessitates solving many independent problems in many fields: image processing, audio processing, networking, artificial intelligence, virtual reality and 3D animation. In designing such a system we divide it into modules, each module being a logical separate unit of the entire system. These modules are: 3D face reconstruction, animation, facial expression recognition, audio processing, interpretation and response generation, and audio-visual synchronization.

The following sub-sections describe the different modules of the system.

3.1 3D Face reconstruction

The 3D face model is constructed from a generic/canonical 3D face using two orthogonal photographs, typically a front and a side view. The process is also referred to as model-fitting as it involves transforming the generic face model to the specific face of the real person. First, we prepare two 2D templates containing the feature points from the generic face – these points characterize the shape and morphology of the face—for each orthogonal view. If the two views of the person to be represented have different heights and sizes, a process of normalization is required. The 2D templates are then matched to the corresponding features on the input images[3]. This employs structured discrete snakes to extract the profile and hair outline, and filtering methods for the other features like the eyes, nose, chin, etc. The 3D coordinates are computed using a combination of the two sets of 2D coordinates. This provides the set of target positions of the feature points. The generic non-feature points are modified using a deformation process called Dirichlet Free Form Deformation (DFFD)[4]. DFFD is a generalized method for free form deformation (FFD)[5] that combines traditional FFD with scattered data interpolation methods based on Delaunay/Dirichlet diagrams. DFFD imposes no constraint on the topology of the control lattice. Control points can be specified anywhere in the space. We can then perform model-fitting using a set of feature points defined on the surface of the generic face as the DFFD control points[6]. For realistic rendering we use texture mapping to reproduce the small details of facial features which may not show up in the gross model fitting. Figure 5 shows the different steps for constructing a 3D face model for the clone.

3.2 Animation

A face model is an irregular structure defined as a polygonal mesh. The face is decomposed into regions where muscular activity is simulated using rational free form deformations [7]. As model fitting transforms the generic face without changing the underlying structure, the resulting new face can be animated. Animation can be controlled on several levels. On the lowest level we use a set of 65 minimal perceptible actions (MPAs) related to the muscle movements. Each MPA is a basic building block for a facial motion parameter that controls a visible movement of a facial feature (such as raising an eyebrow or closing the eyes). This set of 65 MPAs allows construction of practically any expression and phoneme. On a higher level, phonemes and facial expressions are used, and at the highest level, animation is controlled by a script containing speech and emotions with their duration and synchronization. Depending on the type of application and input, different levels of animation control can be utilized. Figure 6 shows these levels.

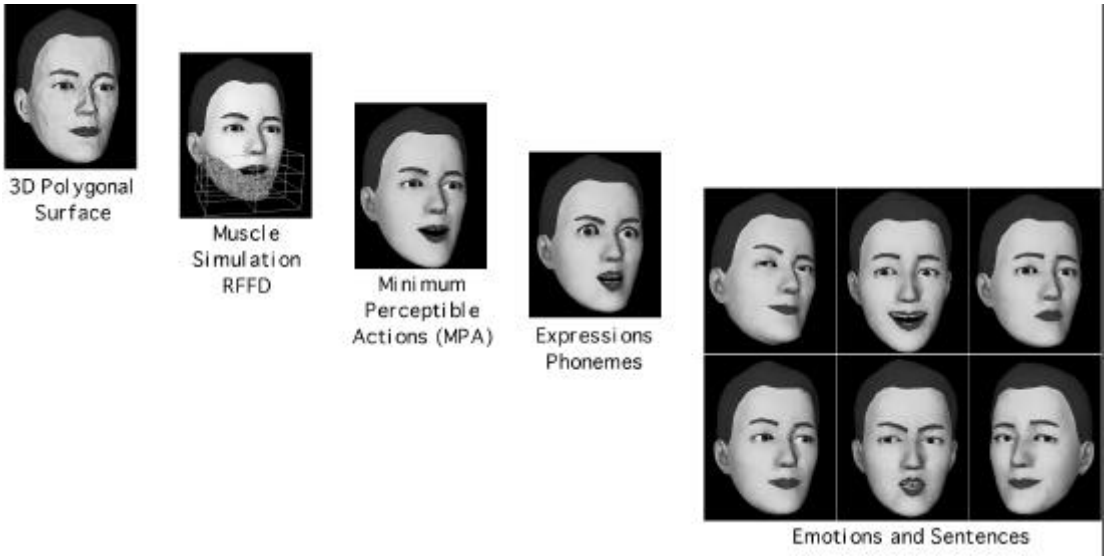

Figure 6: Different levels of animation control.

3.3 Facial expression recognition

Accurate recognition of facial expression from a sequence of images is complex. The difficulty is greatly increased when the task is to be done in real time. In trying to recognize facial expression with a reasonable degree of accuracy and reliability in real time, a few simplifications are inevitable. We focus our attention on only a few facial features for detection and tracking. The method relies on a "soft" mask which is a set of points defined on the frontal face image. During the initialization step, the mask can be interactively adjusted to the image of the real person, permitting detailed measurements of facial features[8]. Other information such as a color sample of the skin, background and hair, etc. are also stored. The feature detection method is based on color sample identification and edge detection. The feature points used for facial expression recognition are concentrated around the mouth, the neck, the eyes, the eyebrows, and the hair outline. The data extracted from previous frame are used only for features which are easy to track (e.g., the neck edges), thus avoiding the accumulation of error. In order to reproduce the corresponding movements on the virtual face a mapping is carried out from the tracked features to the appropriate MPAs, the basic motion parameters for facial animation. This allows us to mimic the facial expression of the real person on their clone. Figure 7 illustrates how the virtual face of the clone is animated using input from the real person's facial expressions.

(a): Working session with the user in front of a CCD camera.

(b): Real time facial expression recognition and animation of the clone's face.

Figure 7: Facial animation of the clone's face.

All this information still does not tell us enough about the mood and/or the emotion of the real person. The person's mood and emotions are important information to be input to the process of formulating the autonomous virtual human's response. We employ a rule-based approach to infer the emotion of the person. These rules are simple mapping of active MPAs to a given emotion, however, these are limited to only a few types of emotion. Another approach to recognizing the basic emotion is to use a neural network with the automatically extracted facial features as input associated to each emotion. We are currently making some experiments for classifying the basic emotions (surprise, disgust, happiness, fear, anger, and sadness) using neural network approach where the input is the extracted data from video and the output is one of the six emotions. Difficulty remains in identifying a blend of basic emotions. This may be partially resolved by identifying the dominant features of the basic emotion depicted and masking the others.

4. Audio processing

The processing of the audio signal (which contains most of the dialog content) to text is non-trivial. Recently, some results have been reported in speaker-dependent restrained contextual vocabulary for continuous speech (90% success rate)[9]. There are, however, many techniques for speech recognition. Typically they involve the following processes: digital sampling of speech, acoustic signal processing (generally using spectral analysis), recognition of phonemes, groups of phonemes and then words (hidden Markov modeling systems are currently the most popular; basic syntactic knowledge of the language may aid the recognition

process). However, these techniques are time consuming and not appropriate for real-time applications. Recent developments suggest that recognition is possible in real time, when used in a simplified context[10] [11]. Audio analysis is essential if the semantics of the input speech is required before formulating a response.

Text-to-speech, or speech synthesis, is another important part of audio processing. The usual way is to split textual speech into small units, generally phonemes, the smallest meaningful linguistic unit. Each phoneme has a corresponding audio signal. It is no easy matter, however, to combine them to produce a fluent speech. One commonly employed solution is to use diphones, instead of just phonemes, which contain the transitions between pairs of phonemes. This squares the number of the elements to be processed in the database but improves the quality considerably. Inflections corresponding to punctuation are added to generate a more human-like voice.

To animate the face using audio input, the audio phonemes are mapped to their corresponding visual output, called the viseme. Since the viseme is defined as set of MPAs, as previously mentioned, it can be applied to any face. Figure 8 shows the same viseme on two different face models.

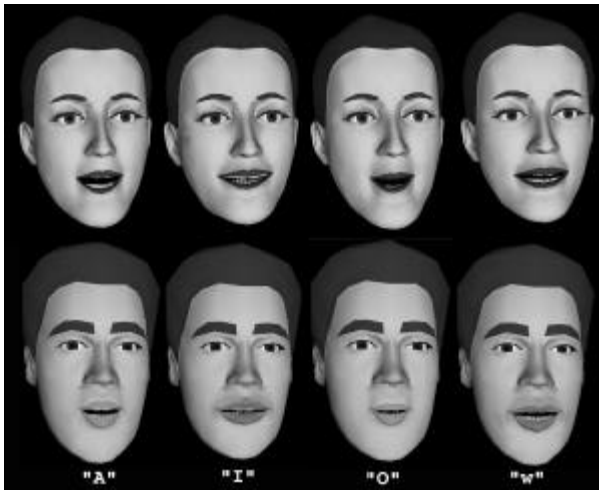

Figure 8: Animation using audio input.

5. Interpretation and response generation

For there to be virtual dialog between the clone and an autonomous virtual human, the autonomous virtual human should be able to "understand" the speech and the emotions of the clone. This requires the addition of an intelligent module to act as the "brain" of the autonomous participant. The information it has to process is composed of the text of the clone's speech and its emotions inferred from facial expressions. The analysis may involve techniques of natural language processing (see Figure 9).

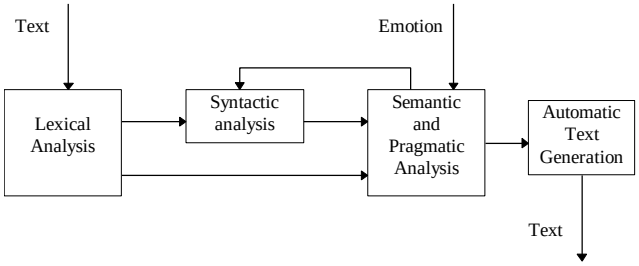

Figure 9. Natural language processing steps.

For each word, the lexical analysis retrieves information stored in a lexicon and then the syntactic analysis relies on a set of grammatical rules to parse each sentence, in order to determine the relations among the various groups of words. The semantic analysis infers the meaning of the sentence and also generates the output used in the automatic response generation. This step can be bolstered by pragmatic analysis of real-world states and knowledge, exemplified in our case by the emotions conveyed by the clone.

Our simplified prototype is an automaton where the final state after the syntactic and semantic analysis is one of the responses in a database available to the autonomous virtual human. This database contains a number of pre-defined utterances, each associated with an emotional state. The current system has limited intelligence, however, this is being extended and elaborated including complex knowledge analysis and treatment.

5.1 Audio visual synchronisation

To produce bimodal output, where the virtual human exhibits a variety of facial expressions while speaking, audio and visual output must be synchronized. To synchronize the sound with the animation, the sound stream is stored in an audio buffer, and the animation, in terms of MPAs, is stacked in an MPA buffer. An MPA synchronizer controls the trigger to both buffers. At present for text to phoneme we are using Festival Speech Synthesis System from University of Edinburgh, UK, and for phoneme to speech (synthetic voice) we use MBROLA from Faculté Polytechnique de Mons, Belgium. Both are public domain software.

5.2 Network issues

From the networking point of view, the virtual dialog system acts like a simple client-server, whose architecture is that of an ongoing networked collaborative virtual environment: Virtual Life NETwork (VLNET)[12]. The clients include the user, as represented by the participation of their clone, as well as one or more external (non-active) viewers.

To reduce bandwidth, only animation or reconstruction parameters of the clone are transmitted to the server. This means that all the video and sound processing has to be computed by the client. The server returns the parameters of the scene to the clients. These include the deformation of objects in the scene, the parameters for speech synthesis of the autonomous virtual human and the compressed speech of the clients and/or clones. The client is free to select a point of view of the scene, with the default being determined by the position of the clone. The client has to reconstruct the view according to the parameters given by the server, decompress the audio and synthesize the speech from the phonemes. Figure 10 shows interaction of the clone with another virtual human while playing chess through network when seen by a third viewer.

Figure 10: Interaction between virtual humans through network.

5.3 Complete pipeline

Figure 11 gives the overview of the complete pipeline showing the data flow among the various modules. The input can be in various forms or media; here we are primarily concerned with video and audio inputs. However, this can be extended for other input such as 3D trackers for body gestures and positions.

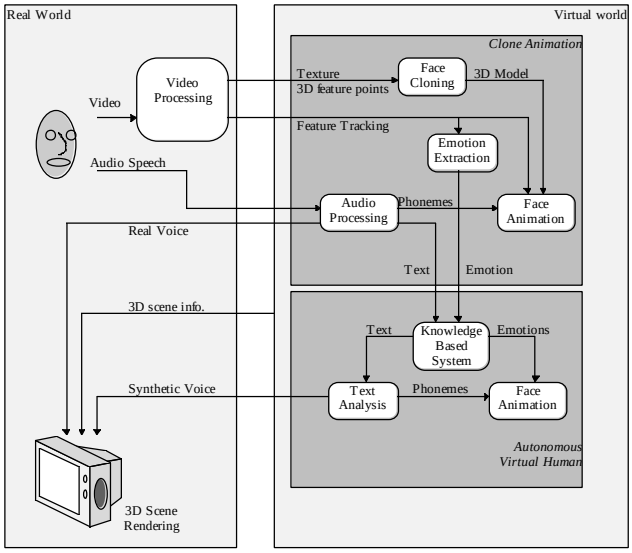

Figure 11. Data flow of the virtual dialogue system.

The audio signal is analyzed to extract the speech content (text) and the phonemes composing this text. In order to synchronize with the animation, we need the onset and duration of each phoneme extracted. The data from the video signal is processed to extract all the visual information about the user, including 3D feature points for modeling the clone face and tracking features for animation.

The autonomous virtual human, a separate entity, receives text and emotions as input. This is processed by the knowledge based module, which then also provides the output response, containing text and emotion. The text is processed to generate temporized phonemes for generating facial deformations as well as synthetic speech. The voice is synchronized with the face deformation, which also accounts for the emotions. Summarizing, the output of the system includes the virtual clone reproducing the motion and speech of the real person, and the autonomous virtual human communicating with the real person, as represented by the clone, via both speech and emotion-conveying expressions.

Figure 11 shows the interaction between only one real face (via clone) and one virtual face, but it can be extended to multi-clone and multi-virtual face communication. This system proves that in principle it is possible, at least in a limited context, to capture and track people's face shapes and movements, recognize and interpret their facial expressions, and produce a virtual dialog, all in real time.

6. Standardisation for SNHC

SNHC (Synthetic Natural Hybrid Coding) is a subgroup of MPEG-4 that is devising an efficient coding for graphics models and compressed transmission of their animation parameters specific to the model type [13]. The University of Geneva is making a contribution to the group as a working partner in VIDAS, a European project formulating a standard set of parameters for representing the human body and the face. For faces, the Facial Definition Parameter set (FDP) and the Facial Animation Parameter set (FAP) are designed to encode facial shape and texture, as well as animation of faces reproducing expressions, emotions and speech pronunciation.

FAP is based on the study of minimal facial actions (like the MPAs in our system) and are closely related to muscle actions. They represent a complete set of basic facial actions and allow the representation of most natural facial expressions. The lips are well defined to take into account inner and outer contours. Exaggerated FAP values permit actions that are not normally possible for humans, but could be desirable for cartoon-like characters.

All parameters involving motion are expressed in terms of the Facial Animation Parameter Units (FAPU). They correspond to fractions of distances between key facial features (e.g. distance between the eyes). The fractional units are chosen to ensure a sufficient degree of precision.

The parameter set contains three high-level parameters. The viseme parameter allows direct rendering of visemes on the face without the intermediary of other parameters, as well as enhancing the result by applying other parameters, thus insuring the correct rendering of visemes. The current list of visemes is not intended to be exhaustive. Similarly, the expression parameter allows the definition of high-level facial expressions.

The FDPs are used to customize a given face model to a particular face. They contain: 3D feature points (e.g. mouth corners and contours, eye corners, eyebrow ends, etc.), 3D mesh (with texture coordinates if appropriate) (optional), texture image (optional) and other (hair, glasses, age, gender) (optional).

7. Conclusion and future work

Providing the computer with the ability to engage in face to virtual face communication - an effortless and effective interaction among real-world people - offers a step toward a new relationship between humans and machines. This paper describes our system, with its different components, which allows real-time interaction and communication between a real person represented by a cloned face and an autonomous virtual face. The system provides an insight into the various problems embodied in reconstructing a virtual clone capable of reproducing the shape and movements of the real person's face. It also includes the syntactic and semantic analysis of the message conveyed by the clone through his/her facial expressions and speech, which can then be used for provoking a credible response from the autonomous virtual human. This communication consists of both verbalizations and non-verbal facial movements synchronized with the audio voice.

We have shown through our system that in a simple situation, it is possible to gather considerable perceptual intelligence by extracting visual and aural information from a face. This knowledge can be exploited in many applications: natural and intelligent human-machine interfaces, virtual collaborative work, virtual learning and teaching, virtual studios, etc. It is not too far-fetched to imagine a situation, real or contrived, where a fully cloned person (face, body, clothes) interacts in a virtual environment inhabited by other virtual humans, clones or autonomous. However, further research effort is required to realize this in real time with realistic and believable interaction. Figure 12 shows a situation where a virtual clone is hanging out with famous virtual actors in a bar (the image is produced from a non-real-time script-based 3D animation).

Figure 12: A bar scene with virtual humans.

Future work will involve recognition and integration of other body gestures and full body communication among virtual humans. We have done some work in the direction of communication between virtual humans using facial and body gestures[14]. However,

cloning of complete virtual humans including animation still needs a lot of research.

8. Acknowledgements

The research is supported by the Swiss National Foundation for Scientific Research and the European ACTS project VIDeo ASsisted with Audio Coding and Representation (VIDAS). The authors would like to thank the members of MIRALab, in particular Laurence Suhner, Laurent Moccozet, Jean-Claude Moussaly, Igor S Pandzic, Marlene Poizat, and Nabil Sidi-Yagoub.

9. References

[1] Cassell J, Pelachaud C, Badler NI, Steedman M, Achorn B, Becket T, Deouville B, Prevost S, Stone M, "Animated Conversation: Rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents", *Proc. SIGGRAPH '94*, pp. 413-420.

[2] Thorisson Kristinn R (1997), "Layered modular action control for communicative humanoids", *Proc. Computer Animation '97*, IEEE Computer Society, pp. 134-143.

[3] Lee WS, Kalra P, Magnenat Thalmann N (1997), "Model based face reconstruction for animation", *Proc. Multimedia Modeling (MMM)'97* , Singapore (to appear).

[4] Moccozet L, Magnenat Thalmann N (1997), "Dirichlet free-form deformations and their application to hand simulation", *Proc. Computer Animation '97*, IEEE Computer Society, pp. 93-102.

[5] Sederbeg TW, Parry SR (1986), "Free-form deformation of solid geometry models". *Computer Graphics (SIGGRAPH '86)*, 20(4), pp. 151-160, 1986.

[6] Escher M, Magnenat Thalmann N (1997), "Automatic cloning and real-time animation of a human face", *Proc. Computer Animation '97*, IEEE Computer Societ, pp. 58-66.

[7] Kalra P, Mangili A, Magnenat Thalmann N, Thalmann D (1992), "Simulation of facial muscle actions based on rational free form deformations", *Proc. Eurographics '92* , pp. 59-69.

[8] Magnenat Thalmann N, Kalra P, Pandzic IS (1995), "Direct face-to-face communication between real and virtual humans", *International Journal of Information Technology* , Vol. 1, No. 2, pp. 145-157.

[9] Philips Speech Processing, URL address *http://www.speech.be.philips.com*

[10] Pure Speech, URL address *http://www.speech.com*

[11] Vocalis, URL address *http://www.vocalis.com*

[12] Capin TK, Pandzic IS, Noser H, Magnenat Thalmann N, Thalmann D (1997), "Virtual human representation and communication in VLNET", *IEEE Computer Graphics and Applications* , March-April 1997, pp. 42-53.

[13] Doenges Peter, Lavagetto Fabio, Ostermann Joern, Pandzic Igor Sunday, Petajan Eric, "MPEG-4: Audio/video and synthetic graphics/audio for real-time, Interactive Media Delivery", *Image Communications Journal* , Vol. 5, No. 4, May 1997.

[14] Kalra P, Becheiraz P, Magnenat Thalmann N, Thalmann D (1997), "Communication between synthetic actors", Chapter in the book: Automated Spoken Dialogues Systems (Ed Luperfoy S), MIT Press, Cambridge, MA (to appear).