Multimodal Human Machine Interactions in Virtual and Augmented Reality

Gérard Chollet[1], Anna Esposito[4], Annie Gentes[1], Patrick Horain[3], Walid Karam[1], Zhenbo Li[3], Catherine Pelachaud[1], Patrick Perrot[12], Dijana Petrovska-Delacrétaz[3], Dianle Zhou[3], and Leila Zouari[1]

[1] CNRS-LTCI TELECOM-ParisTech, 46 rue Barrault, 75634 Paris - France,
[2] Institut de Recherche Criminelle de la Gendarmerie Nationale (IRCGN), Rosny sous Bois - France,
[3] TELECOM & Management SudParis, 9 rue Charles Fourier, Evry - France
[4] Second University of Naples, Dept. of Psycology, and IIASS, Italy
{gerard.chollet,annie.gentes,walid.karam,catherine.pelachaud,patrick.perrot,leila.zouari}@telecom-paristech.fr
{patrick.horain,dijana.petrovska,dianle.zhou,zhenbo.li}@it-sudparis.eu
{iiass.annaesp}@tin.it

Abstract. Virtual worlds are developing rapidly over the Internet. They are visited by avatars and staffed with Embodied Conversational Agents (ECAs). An avatar is a representation of a physical person. Each person controls one or several avatars and usually receives feedback from the virtual world on an audio-visual display. Ideally, all senses should be used to feel fully embedded in a virtual world. Sound, vision and sometimes touch are the available modalities. This paper reviews the technological developments which enable audio-visual interactions in virtual and augmented reality worlds. Emphasis is placed on speech and gesture interfaces, including talking face analysis and synthesis.

Key words: Human Machine Interactions (HMI), Multimodality, Speech, Face, Gesture, Virtual Words

1 Introduction

Humans communicate through a gestalt of actions which involve much more than speech. Facial expressions, head, body and arm movements (grouped under the name of gestures) all contribute to the communicative act. They support (through different channels) the speaker's communicative goal, allowing the speaker to add a variety of other information to his/her messages including (but not limited to) his/her psychological state, attitude, etc. The complexity of the communicative act should be taken into account in human-computer interaction research aiming at modeling and improving such interaction by developing user-friendly applications which should simplify and enrich the end user's ability to interact with automatic systems. Psycholinguistic studies [26, 25, 43, 52, 56] have confirmed the complementary nature of verbal and nonverbal aspects in

human expressions, demonstrating how visual information processing integrates and supports the comprehension of messages. In the light of these results, several research works on mutual contribution of speech and gesture in communication and on their characteristics have being carried out also in the field of Human Machine Interaction (HMI). Such studies are mainly devoted to implement synchronization between speech, facial, and body movements in human communicative expressions [30, 35, 41, 45, 55, 61, 73]. Some models include characteristics of emotional expressions [24, 33, 56] as well as speech and gesture modeling, multimodal recognition and synthesis (virtual agents) of facial expressions, head, hands movements and body postures [1, 46, 75, 32, 60, 63].

Psycholinguistic studies on human communication have also shown that humans convey meaning by a set of nonlexical expressions carrying specific communicative values such as feedback and regulation of the interaction. Preliminary studies have observed that such nonlexical speech events are highly communicative and show gestural correlates [26, 23] stressing the need for their mathematical modeling so as to implement more natural "talking heads" and/or "talking faces".

The above research arguments have led to the development of virtual worlds and virtual realities, where software characters and intelligent devices interact with humans through the plurality of communication modes and/or develop mirror artificial lives in a 3D virtual world created by avatars that perceive, reason and react showing cognitive workloads, consciousness and emerging behaviors. The word "avatar" originates from the Hindu culture, meaning a god's coming to earth in bodily form [19]. Nowadays, it is an embodiment, as of a quality in a person, and it is used in virtual worlds to refer to the graphical representation of a human being in virtual environments. Its appearance, behavior and communication skills are of great importance in virtual worlds.

The present paper is situated in such a context as it reports on the recent research efforts for the development of new models, methods, and communication paradigms to improve information and communication technologies. The rest of the paper is organized as follows. In Sect. 2 issues related to designing 3D applications are presented. Section 3 summarizes the architecture and standards related to Embodied Conversational Agents (ECAs). In Sect. 4 examples of technologies that have been developed to endow systems with acoustic and visual perceptive and generative capabilities are given. They include speech, face and gesture processing. Conclusions and perspectives are given in Sect. 5.

2 Designing 3D Applications: Challenges, Co-design Methodology and Results

3D worlds host a number of activities, organizations, representations and purposes. Today, Second Life[5] is used as a showroom for products or channel of information. People can: visit stores and manipulate virtual objects, organize

[5] www.secondlife.com

or join meetings, be interviewed for new jobs by companies such as L'Oreal, or BNP Paribas [29]. The amount of interactions and creations of artefacts testifies to the success of such shared environments [8]. But one may wonder what are the main venues of innovation, apart from increasing the amount of transactions, the number of participants and solving technical issues such as bandwith, persistency of data, etc. To provide an answer, one has to enquire not only into the actual uses and organizations of these 3D environments, but also into other 3D applications and domains of research and to figure what is relevant to the development of new interactions.

Our hypothesis was to associate three axes of research: face and gesture processing, speech recognition, and embodied conversational agents. Our interest is to study their potential in creating significant interactions within 3D environments. The method focuses on combining these technologies to create a situation of interaction. Such creation is based on the analysis of 3D practices, design and tools as well as a sociological study of 3D environments. The purpose is to design a convincing situation and interface so that end-users may actually enjoy an enhanced 3D experience. The design of a demonstrator is to ensure that testers will be able to assess the results and the potential of such a convergence.

3D application research focuses on three main issues. How to: make avatars and contexts more realistic and highly expressive without relying on post production; automate and delegate some behaviors to avatars; and improve interactions in 3D environments. Each area of research is highly challenging as users are expecting very sophisticated representations of characters and settings in relation to our culture of animation movies or videogames. While a certain latitude is acceptable in research circles that allow for basic aesthetics features, it is difficult to design a proper interaction that end-users won't find disappointing. To face this challenge, we followed a co-design methodology which takes into consideration the strong points but also the limits of these technologies as well as some anthropological specificities of virtual worlds.

Virtual worlds are places of creativity where people draw, stage, and invent stories. They offer new modalities to express oneself and communicate with others that can compete or complete traditional media. Primarily, virtual worlds offer a place for people to play with different aspects of their personality. As was analyzed by Sherry Turkle [78] as early as in the 80s, when the first Moos and Muds appeared, the appeal of such systems is in exploring a different self and testing it with other online users. We decided to rely on this sociological and psychological characteristics to design a demonstrator that offers a situation of reflexive learning. The system is a mirror to the user's activity, helping her to analyze and correct her behavior in a situation of recruitment. We also wanted to explore the multidimensional potential of 3D environments including a diversity of information and access: the applicant CV, the online information on the company that is recruiting, the video of the interaction between the avatar and the Embodied Conversational Agent (ECA) that can be replayed. Our hypothesis is that the future of 3D applications is a seamless staging of 2D and 3D elements,

texts, images and videos, so that the users can benefit from all ofthem without having to open a new system for each.

The choice of this learning program presents other interesting features. Though the interviews can be varied, they rely on a basic vocabulary centred on expertise, studies, work experience, etc. of the candidate. In terms of acquisition, the database can easily be augmented by the vocabulary of the company—that can be found on its website—and by the CV of the applicant that he can feed the system with. We therefore limit the scope of what the speech recognition system will have to manage, increasing its reliability. The dialog form is also relatively easy to anticipate: greetings, presentation of each other, turns of speech, objectives of the interview, can be defined that will allow for the modeling of the ECA responses to the user. The number of gestures and face expressions is also limited and therefore can be more accurate.

The analysis of the technical characteristics, of the anthropological features of 3D worlds, and of the formal issues (such as access, format, navigation, interface, etc.) provide the theoretical framework for new applications. Here, co-designing means that we are taking into consideration each of these elements, elements that we reformulate and organize in a coherent context. This methodology ties together the inner workings of the technology and its outer workings, including a definition of its original aesthetics.

3 Architecture and Standards

Several multimodal real-time ECA architectures have been implemented [11, 46, 13]. MAX (the Multimodal Assembly eXpert) [46] is a real-time system where a 3D agent evolves in a Cave Automatic Virtual Environment (CAVE). The system is composed of three high level modules: *Perceive*, *Reason* and *Act*. Such an architecture allows the agent to perceive the world it is placed in and to react to any event either cognitively if processing time permits it or reactively. BEAT (the Behavior Expression Animation Toolkit) takes a text as input and provides an animation as output [13]. The architecture is composed of three modules, language tagging, behavior generation and behavior scheduling. The text is sent to an external Text-To-Speech (TTS) system that returns the list of phonemes and their duration. This last information with the behavior description is sent to the animation module. Multi-agent architectures are often used in ECA systems (Open Agent Architecture [14], Adaptive Agent Architecture or even Psyclone [74]). In particular, white board architecture is used to ensure the exchange of messages between the modules (eg. Psyclone [74]). It allows modules and applications to interact together, even if they are running on separate machines connected through TCP/IP.

Situated SAIBA [80] (Situation, Agent, Intention, Behavior, Animation) is a current international effort to define a unified platform for ECAs. SAIBA aims at developing modular and distributable architecture. Its goal is to allow for rapid integration within existing systems. This framework is composed of three main steps: (1) communicative intent planning, (2) multimodal behavior planning, and

(3) behavior realization (see Figure 1). The first step outputs what the agent intends to communicate and in which emotional state it potentially is. The second step computes through what verbal and nonverbal means the agent will communicate its intent and emotion. Finally the last steps instantiate these multimodal behaviors in audio and animation files. The interaction between these different steps is done via two markup languages. Between the first two steps, there is the Functional Markup Language (FML) that describes the communicative intent; then the last two steps are connected via the Behavior Markup Language (BML). The animation of the agent is specified through the Facial Animation Parameters (FAPs) and the Body Animation Parameters (BAPs) of the MPEG-4 standard. Lately Thiebaux and colleagues [73] have developed a real-time architecture to implement the *Behavior Realizer* of the SAIBA framework.

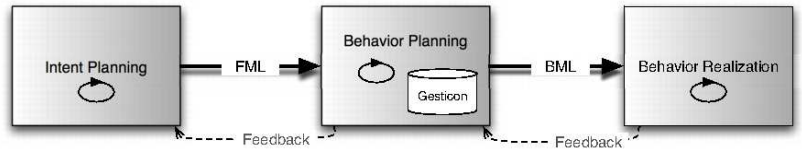

Fig. 1: The Situation, Agent, Intention, Behavior, Animation (SAIBA) framework [80]

4 Multimodal Interaction

Communication is multimodal by essence. It involves not only speech but also nonverbal signals. Voice intonation, facial expressions, arm and hand gestures, gaze, etc, are all communication signals. Moreover communication does not involve separate processes of perception (by the listener) or of generation (by the speaker). Both interactants, speaker and listener, send and perceive signals from the others. Interaction involves a cocontinuous exchange of messages where interactants adapt to each other. Multimodal interaction systems ought to include perception, generation and adaptation capabilities [46, 75]. In this section, we present technologies that have been developed to endow systems with acoustic and visual perceptive and generative capabilities.

4.1 Speech Processing

Recognition: In order to make a vocal conversation between an ECA and a human user, the ECA must be able to recognize and understand as much as possible what the human says. Hence, a speech recognition system is needed.

Speech recognition consists in converting the acoustic signal into a set of words (sentences). In the framework of a vocal human machine interaction, these

words serve as an input to further linguistic processing in order to achieve speech understanding.

Recent works on speech recognition are based on statistical modeling [27, 57, 86, 9, 37]: given a sequence of observations $O = o_1, o_2, ..o_T$ (the speech samples), the recognition system looks for the most probable set of words $M^{'} = m_1, m_2, .., m_N$ (knowing the observations):

$$M^{'} = argmax\ p(M/O) = argmax\ p(O/M) \times p(M) \tag{1}$$

where :

- M are the words of the application dictionnary.
- $p(O/M)$, the probability of the observation, given the words, is given by an acoustic model. The Hidden Markov Models (HMMs) are widely used for acoustic modeling and the modeled units are often phonemes, syllables or words.
- $p(M)$, the probability of the words, is given by a linguistic model. A linguistic model aims to restrict the combination of words.

Therefore the global architecture of a speech recognition system is as depicted in the following figure.

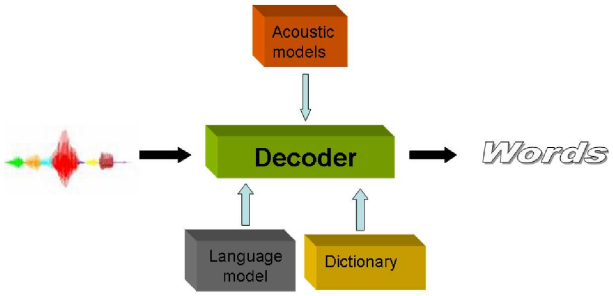

Fig. 2: Architecture of a speech recognition system

Performance of speech recognition systems is often described in terms of Word Error Rate (WER).

$$WER = 100 \times \frac{substitutions\ +\ omissions\ +\ insertions}{number\ of\ correct\ words} \tag{2}$$

The WER depends on different parameters [57]:

- The speaking mode: isolated words are easier to recognize than continuous speech.
- The speaking style: usually spontaneous speech contains disfluencies and is more difficult to recognize than read speech.

- The size of the vocabulary: when the vocabulary is large, there are more confusions between words.
- The speaker dependency: when the system is speaker dependent, adaptation techniques can be used to improve the modeling.
- Recording conditions: noisy or clean, wide- or telephonic-band speech, etc.

Synthesis: The ECA should be able to talk and interact meaningfully with humans using speech. Most developments in speech synthesis have focused on text to speech systems [21]. A text-to-speech synthesizer is a system that should be able to read any text introduced in the computer. To develop such a system, three main approaches have been proposed:

- Articulatory synthesis: in the early days of synthesis, researchers simulated human speech production mechanisms using articulatory models. Speech is created by simulating the flow of air through the representation of the vocal tract.
- Acoustic synthesis-by-rule: later, text processing was modeled by a set of rules (phonetic theories and acoustic analysis). The resulting technology is referred to as speech synthesis-by-rule. Most of synthesis-by-rule progress has been system dependent due to detailed rules and finely tuned parameters. As a consequence, the expert knowledge can be very hard to reproduce in other systems.
- Concatenative synthesis: speech is produced by combining fragments of prerecorded speech. Currently, great progress has been made using large-corpus concatenative approaches. Many systems output speech indistinguishable from that of human speakers [17].

The most important characteristics of a speech synthesis system are naturalness and intelligibility. While work on naturalness continues, the next step in the improvement of intelligibility and conviviality of the text-to-speech systems concern their expressiveness, referred as expressive synthesis.

Voice Transformation: Speaker transformation (also referred to as voice transformation, voice conversion, speaker forgery, or speaker adaptation) is the process of altering an utterance from a speaker (impostor) to make it sound as if it were articulated by a target speaker (client). Such transformation can be effectively used by an avatar to impersonate a real human and converse with an ECA.

Speaker transformation techniques [28, 39, 85, 40, 2, 7, 77, 62, 16] might involve modifications of different aspects of the speech signal that carries the speaker's identity. We can cite different methods. There are methods based on the coarse spectral structure associated with different phones in the speech signal [39]. Another techniques use the excitation function (the "fine" spectral details, such as [85]). Prosodic features representing aspects of the speech that occur over timescales larger than the individual phonemes can also be used, as well as the

"Mannerisms" such as particular word choice or preferred phrases, or all kinds of other high-level behavioral characteristics. The formant structure and the vocal tract are represented by the overall spectral envelope shape of the signal, and thus are major features to be considered in voice transformation [40].

The voice transformation techniques proposed in the literature can be classified as text-dependent or text-independent methods. In text-dependent methods, training procedures are based on parallel corpora, i.e., training data have the source and the target speakers uttering the same text. Such methods include vector quantization [2, 7], linear transformation [38, 84], formant transformation [77], vocal tract length normalization (VTLN) [71], and prosodic transformation [7]. In text-independent voice conversion techniques, the system is trained with source and target speakers uttering different texts. Text-independent techniques include VTLN [71], maximum likelihood adaptation and statistical techniques, unit selection, and client memory indexing [62, 16]. The analysis part of a voice conversion algorithm focuses on the extraction of the speaker's identity. Next, a transformation function is estimated. At last, a synthesis step is achieved to replace the source speaker characteristics by those of the target speaker.

Consider a sequence of a spectral vectors pronounced by the source speaker, $X_s = [x_1, x_2, .. \ x_n]$, and a sequence pronounced by the target speaker composed of the same words, $Y_s = [y_1, y_2, .. y_n]$. Voice conversion is based on the calculation of a conversion function F that minimizes the mean square error, $E_{mse} = E[||y - F(x)||^2]$, where E is the expectation.

Two steps are needed to build a conversion system: a training step and a conversion step. In the training phase, speech samples from the source and the target speaker are analyzed to extract the main features. Then these features are time aligned and a conversion function is estimated to map the source and the target characteristics (Fig. 3). The aim of the conversion is to apply the estimated transformation rule to an original speech pronounced by the source speaker. The new utterance sounds like the same speech pronounced by the target speaker, i.e. pronounced by replacing the source characteristics by those of the target voice. The last step is the re-synthesis of the signal to reconstruct the source speech voice (Fig. 4). Other approaches have also been proposed in [62].

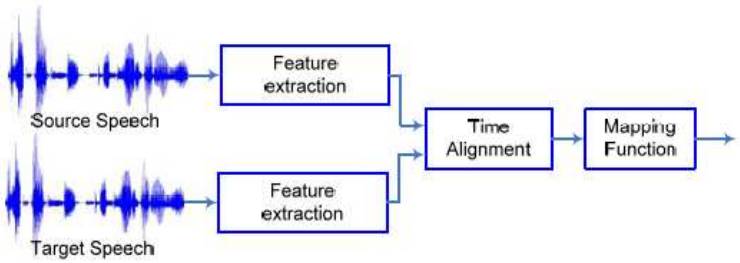

Fig. 3: Training step

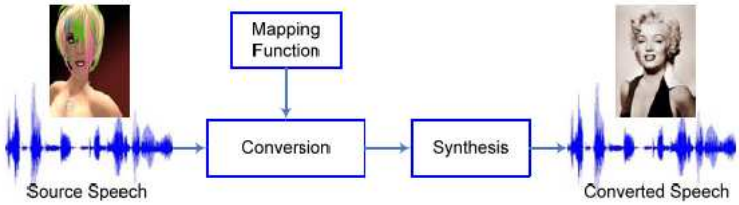

Fig. 4: Conversion step

A way to evaluate the conversion is to measure the displacement of the impostors distribution towards the client distribution. Voice conversion can be effectively used by an avatar to impersonate a real human and hide his identity in a conversation with an ECA. This technique is complementary with face transformation in the creation of an avatar that mimics the voice and face of a real human target.

Speaker Verification: In a virtual word, speaker verification could be used to authenticate an avatar, except if voice conversion is available. It could be necessary to authenticate an avatar before starting a dialog, for a bank transfer for instance, but also just to avoid embarrassing confusions in the communication. The speaker verification step could be a preliminary to all kind of communication in a virtual word, where the identity is important. In this paragraph the state of the art of speaker verification is summarized. A more detailed overview is available in [31].

Speaker verification consists in verifying a person claimed identity. Speech actually conveys many characteristics providing discriminative information about the speaker. These characteristics are divided in "high level" and "low level" attributes. The first set of features (high level) includes prosody, phonetic information, pronunciation defaults. The low level features are the ones related to the physical structure of the vocal apparatus. In most cases, automatic speaker recognition system are based on the low level features such as the MFCC (Mel Frequency Cepstral Coefficient) and their derivatives. Some alternative representations are also used, such as the LPC (Linear Predicting Coding) based cepstral features and their derivatives. This step of feature extraction consists in extracting a time sequence of feature vectors representative of the temporal evolution of speech spectral characteristics.

After the parameterization step, generative models including GMMs (Gaussian Mixture Models) and HMMs (Hidden Markov Models) are built to represent the distribution of the speaker's feature vectors. GMMs are defined by a set of parameters. During the training phase these parameters are iteratively estimated using the Expectation Maximization algorithm. Adaptation techniques are usually useful because of the lack of data. The aim of this step is to adapt the parameters of the background model with the speaker's enrollment speech data.

Some other techniques can be used based on discriminative algorithms, such as MLP (Multilayer Perceptrons) or SVM (Support Vector Machines).

The principle of a speaker verification task is to take a decision whether the speech data collected during the test phase, belongs to the claimed model or not. Given a speech segment X and a claimed identity S the speaker verification system should choose one of the following hypothesis:

H_S**:** X is pronounced by S
$H_{\bar{S}}$**:** X is **not** pronounced by S

The decision between the two hypothesis is usually based on a likelihood ratio given by:

$$\Lambda(X) = \frac{p(X|H_S)}{p(X|H_{\bar{S}})} \begin{cases} > \Theta & accept\ H_S \\ < \Theta & accept\ H_{\bar{S}} \end{cases} \tag{3}$$

where $p(X|H_S)$ and $p(X|H_{\bar{S}})$ are the probability density functions (called also likelihoods) associated with the speaker S and non-speakers $\bar{S}$, respectively. Θ is the threshold to accept or reject H_S.

Based on this decision, it is possible to claim a probable identity and to verify if the person behind the avatar is the claimed person.

4.2 Face Processing

Face-to-face communication implies visually perceiving the face of engaged people. When interacting in a virtual environment, face perception should be supported through that environment. This requires that the machine will have the capability to see human faces, namely to detect, track and recognize faces, and to render virtual actors with expressive animated faces. In addition, face metamorphosis allows controlling one's virtual appearance.

Face Detection and Tracking Capturing head pose and facial gesture is a crucial task in computer vision applications such as human-computer interaction, biometrics, etc. It is a challenge because of the variability of facial appearances within a video sequence. This variability is due to changes in head pose (especially out-of-plane rotations), facial expressions, lighting, occlusions, or a combination of all of them.

Face detection: Capturing head pose and facial gesture is a crucial task in computer vision applications such as human-computer interaction, biometrics, etc. It is a challenge because of the above mentioned variabilities of facial appearance within a video sequence.

Feature-based tracking: Face motion and deformation between successive frames of a video sequence can be determined by tracking face features. These are highly discriminative areas with large spatial variations such as eyes, nostrils, or mouth corners to be identified and tracked from frame to frame. Contour of

Fig. 5: Face detection using the Adaboost algorithm

features can be tracked with active contours, or "snakes", that obey internal and external forces. Internal energy terms account for the shape of the feature and smoothness of the contour while the external energy attracts the snake towards feature contours in the image [72]. Face features (eyes, mouth) can be described with sets of wavelet components and linked together in an elastic graph [81].

Appearance-based tracking: An appearance-based tracker matches a model of the entire facial appearance with an input image. The problem of finding the optimal parameters is a high dimensional search problem. Active Appearance Models (AAMs) introduced by Cootes et al. [18] are joint statistical models of the shape and texture of an object (e.g., a face), combined with a gradient descent algorithm for adapting the model to an image. Their ability to vary their shape and appearance to match a person's face relies on an extensive learning of the face variations with respect to identity, expression, pose and/or illumination. Matthews and Baker [51] introduce an inverse compositional algorithm to accelerate image alignment for AAM. Xiao et al. [83] show how to constrain the 2D AAM search with valid 3D model modes. Ahlberg has extended AAM to a 3D face model [4]. Dornaika and Ahlberg [20, 5] present a two steps algorithm in which they apply a global RANdom SAmple Consensus (RANSAC) method to estimate the rigid motion of the head in-between video images and a local optimization scheme to estimate the facial animation parameters.

Synthesis The expressions of the six prototypical emotions are often used in 3D agents technology [58]. The standard MPEG-4 includes them in its specification [58]. These models are based on the categorical representation of emotion, mainly on Ekman's work [22]. In the EmotionDisc model developed by Ruttkay [67], the six prototypical emotions are spread uniformly on a circle. Each quadrant of the circle is further divided. The closest part to the center corresponds to the highest intensity of the emotion represented in this quadrant. The facial expressions associated to any point of the circle are computed as a bi-linear interpolation between the closest known expression of emotion and the neutral expression. Models of facial expressions based on a dimensional representation of emotions have been proposed [76, 6]. In these works, the facial expression associated to an emotion defined by its 3D coordinate in the emotional space is

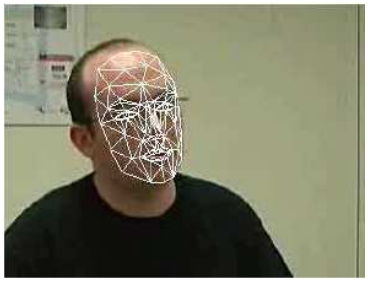

Fig. 6: Real-time face tracking with an Active Appearance Model (video available from http://picoforge.int-evry.fr/projects/myblog3d)

computed as a linear interpolation of the closest prototypical emotions found in the 3D space. A model of complex emotion, ie. emotions arising as a combination of emotions, has been implemented using a fuzzy logic by [10, 55]. Duy Bui [10] elaborated a model to blend the expressions of the six prototypical emotion. Niewiadomski et al. [55] extended this model to work on a larger set of complex emotion types and of emotions. In this model, fuzzy rules are defined for each type of complex emotions. Several types are considered: superposition of emotions, masking an emotion by another one, inhibition viewed as masking an emotion by a "neutral" one, etc.

Face Transformation Different transformations could be considered when a change from one facial appearance to another is requested. They depend on the goal that has to be reached and the available methods that could be used. Facial transformation techniques could be useful for applications, such as customization of an avatar or face transformations methods used for criminalistic purposes.

The variety of representations that an avatar can have are actually limited by the technological possibilities of computer graphics and the imagination of end-users and programmers. End-users should be able to customize their own avatars in order to differentiate them from those of other users. It is important that they have at their disposal the tools necessary for this customization. Currently, systems employ a library of generic bodies, body parts and accessories for avatar creation. The face is one of the most easily recognizable, and expressive features of the human body and it seems natural to attempt to integrate it into the avatar itself. There are some attempts to create personalized avatars, but doing so for 3D facial representations in an automatic way is not a widespread usage. In this paragraph, three examples of face transformation are presented. The first one is related to 2D face morphing. In the second example, a 3D face is reconstructed, using two 2D images (with frontal and profile views) from an individual. In the third example, it is shown how to create a personalized 3D avatar (the face part), using only one 2D face image of that person.

2D face morphing: is a good way to approach progressively one's own face from an avatar for instance. There are many methods to perform face morphing. All of them require the correspondence of landmark points between the source and the target faces. The landmark points could be found manually or automatically. The method presented in [87], does not require human intervention, using a generic model of a face and evolution strategies to find the control (also called landmark) points in both face images. Automatic landmark methods are also used in [42], where a genetic algorithm is used to extract control points. These control points, generally correspond to the eyes and the mouth (like in [47]), and sometimes the nose position is also needed [42]. Recently Wolberg proposed a novel method that is based on the application of a multi-level free-form deformation [82]. The 2D morphing example presented in this paragraph needs a selection of control points by the user, as well as the number of intermediate images required. morphing is smoother. All operations are made on rectangles rather than on triangles, which decreases the computing times. The morphing process can be defined as follows: make a smooth and fast transformation from a source image to a target one. This process can be divided into three parts: the control points selection, the warping, and the final blending step. The first step of the face morphing is to select the control points such as eyes, the tip of the nose, both extremities of the mouth, and both extremities of the chin (Fig. 7).

a) b)

Fig. 7: 2D face morphing: (**a**) control points selection, and (**b**) division of an image

After the control point selection step, a warping process is necessary. This process consists in moving the control points of the source image to match the ones in the target image. For two corresponding rectangles, one from the source image and the other one from the target image, this transformation can be performed via the linear transformation method. After this step, the source image control points are matched with the target image control points, thus the application of the next step, blending, will produce a smooth morph, without artifacts.

At this point in the face morphing process, all important elements of the source figure are on the same position as that of the elements of the target figure, that is to say that the source eyes are on the same position as the target eyes, etc. That means that all the elements of the source figure may have been

stretched, or shrunk. Then, in order to have the same color for the eyes, the skin, and so on, a color transition has to be performed.

What is interesting in such a case is to morph a face progressively towards another one. Many applications can be imagined, based on this principle like for instance the personalization of an avatar in a virtual world.

3D face reconstruction with two facial images: In addition to the above morphing process, it is interesting to reconstruct a 3D from two photos of an individual [59]. The idea is to reconstruct one's own face for instance in 3D and then to morph the avatar face toward one's face. The aim is to personalize an avatar in order to progressively reveal one's identity. From two photos of the same person (one from

Fig. 8: Example of the two 2D input images, used to reconstruct the 3D model

a frontal view, the other one from a profile view, as shown in Fig. 8), it is possible to reconstruct a 3D face. The principle is to use a generic model in 3D and to adapt it to the deformation of the photo. After a vertical adjustment between the two photos, depth data are extracted from the profile view and width data from the frontal view, by selecting corresponding pairs of points between a photo and the 3D generic model. Thus, a global 3D morphing matrix is computed. This matrix is then applied to the whole generic 3D model to obtain the deformed 3D model (Fig. 9). Geometric and color maps are produced by ray-tracing on the 3D deformed model, with a different grid for each photo. For each coordinate (x, y, z) and the color coordinates, two maps, one for each view, are obtained. They must finally be blended so as to produce a smooth colored 3D model.

Creation of a personalized 3D avatar using one 2D face image: there are some attempts to create personalized 3D face avatars, when only one 2D face image is available. The most important works are those from S. Romdhani et al. [66] related to Morphable Models of Faces. The main idea of morphable faces is to use a statistical (also called generic) 3D face model, in an analysis by synthesis framework in order to synthesize a 3D face model that resembles the face from an unseen input 2D face image. Such a framework is described in Fig. 10. An important issue in this framework is the creation of the generic 3D face model. The idea is to use the 3D a priori knowledge of a set of known 3D human faces in order to reconstruct new 3D models of subjects for whom a 3D face scan is not available.

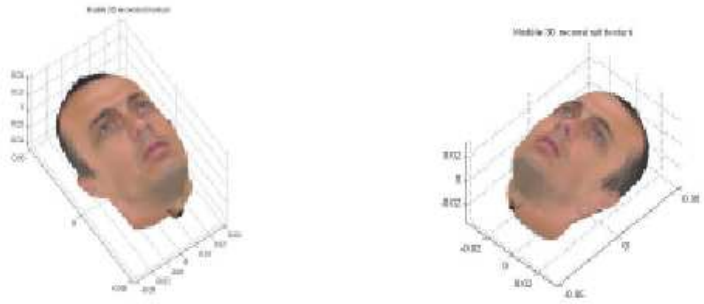

Fig. 9: The reconstructed 3D face, using the two 2D face images, shown in Fig. 8

The example presented in this paragraph is based on the general framework of morphable 3D faces in an analysis by synthesis loop (Fig. 10). There are two main phases in this framework. The first phase is the construction of the generic 3D face model (from an ensemble of 3D scans acquired on real subjects). The second phase is the fitting part, in order to find the fitting parameters of the model that result in a representation as close as possible to the input image.

S. Romdhani et al. [66] use the CyberwareTM laser scanner to acquire full 3D scans of 200 subjects, with shape and texture data in a cylindrical representation. For the example presented in this paragraph, the 3D partial scans from the IV^2 face database (described in [64] and available at [36]) are used. These 3D scans are acquired with the Konica-Minolta Vivid 700 laser sensor. This scanner provides partial 3D information. Three partial face scans (frontal, left and right profile) are employed for the construction of the full IV^2-3D face generic model. Some semi-automatic treatments, such as smoothing, hole filling, hair and ear removal, and noise removal need to be done on the partial scans. The Iterative Closest Point (ICP) algorithm is applied for the registration phase in order to merge the shape partial scans.

The face morphing and 3D reconstruction methods presented above are interesting tools to personalize an avatar in 3D virtual worlds, or to choose different avatars that for instance look like one's favorite actress or actor.

4.3 Gesture Analysis and Synthesis

Gestures are part of human communication, so they should be supported with machine perception and rendering for mediated communication in virtual environments. Two main techniques are used to synthesize gesture: motion capture and key-frame animation. While the first technique produces high quality animation the second one is much more flexible. To animate an ECA procedural animation is often chosen over motion capture technique. The representation of co-verbal gestures in ECAs follows the literature on communicative gesture [43]. A gesture is divided into phases: preparation, pre-hold-stroke, stroke, post-stroke-hold and retraction [43]. Several systems use the approach proposed

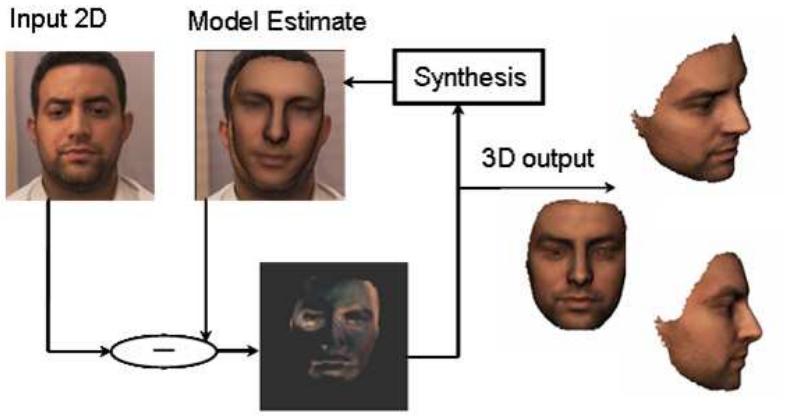

Fig. 10: An example of a personalized 3D face avatar, obtained by the analisis and synthesis framework, using only one input 2D face image of the subject

by Liddell [48] to describe arm position and hand shape for American Sign Language. Each phase of a gesture is defined by the wrist location in the gesture space, palm and finger orientation, finger and thumb shape [46, 32]. An interpolation scheme is defined to smoothly join between arm position for each phase. Languages to describe gesture shape for each phase have been elaborated. Examples are MURML [46], Greta formational parameters [32], BML [80]. Cassell and her colleagues developed a representation of gestures that ensure a generation "on the fly" of speech and gesture, depending on the semantic information both modalities convey [12]. The manner of behavior is tackled by Chi et al. [15] who implement the EMOTE model of adapting agent gestures in order to enhance expressivity of behaviors applying transformation at the level of limbs. Hartmann et al. [32] obtained a set of expressivity parameters, tying behavior modificationinto the synthesis stage of gesturing: Mancini and Pelachaud [50] extended this to include facial expressions and head movements. Kipp et al [44] present a gesture animation system based on statistical models of human speakers gestures. Videos of interviewed people have been manually annotated in terms of gestures types (iconic, deictic, etc.) together with their frequency of occurrence and timing (that is the synchronization between the gesture stroke and the most emphasized syllable of the sentence). The statistics on the speaker's gestures are then used to model the agent's set of preferred gestures (the probabilities of their occurrence is computed from the annotated gesture frequency) and synchronization tendency (for example an agent can perform gesture strokes always synchronized with speech emphasis). Mancini and Pelachaud [50] propose a computation model to create ECAs with distinctive behaviors. ECA is defined by a baseline that states the global tendency of the ECA (modality preferences and overall expressivity values for each modality) and a dynamic line that describes how a given communicative intent and/or an emotional state affect the baseline.

Gesture Modeling for Human Motion Tracking

Motion capture by computer vision: Computer vision has become a user-friendly alternative to sensors attached on user limbs [53]. The interest in this research has been motivated by many potential applications (video surveillance, human computer interaction, games, etc.) and their inherent challenges (different human body proportions, different clothes, partial occlusion, uncontrolled environments, etc.) [65].

Recent research can be divided into two main classes. Model-based (or generative) approaches use a model of the human body and a matching cost function to search the pose that best matches input images. Model-free (or discriminative) approaches try to learn a direct relation between image observation and the pose [68]. Different techniques have been proposed using these approaches, as well as using different information sources (color, edge, texture, motion, etc.) [49], sample-based methods [69], and learning methods [3].

We have developed a prototype system for 3D human motion capture by real-time monocular vision without marker. It works by optimizing a color-based registration of a 3D articulated model of the human body on video sequences [34]. Model candidate postures are compared with images by computing a non-overlapping ratio between the segmented image and the 3D-color model projection. This evaluation function is minimized to get the best registration. Biomechanical limits constrain the registered postures to be morphologically reachable. Statistical dynamic constraints enforce physically realistic movements. This has been implemented to run at video frame rate [70], allowing virtual rendering of captured motion in real time.

Fig. 11: Real-time motion capture from webcams and virtual rendering (video available from http://picoforge.int-evry.fr/projects/myblog3d)

Gesture modeling: Gesture models in low dimensional latent space can be useful as prior knowledge for human motion tracking, especially with monocular vision.

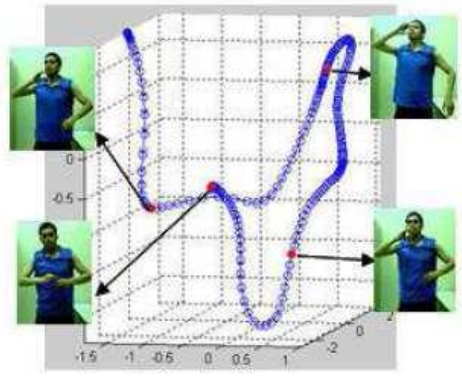

Fig. 12: A GPDM model of drinking. Each point in the 3D latent space corresponds to a human pose.

Furthermore, tracking in a lower dimensional latent space requires fewer particles for particle filtering. Urtasun [79] used GPLVM and GPDM to learn prior models for tracking 3D human walking, so achieving good results even in case of serious occlusion. Moon and Pavlovic [54] have investigated the effect of dynamics in dimensionality reduction problems on human motion tracking. Gesture models in low dimensional latent space may receive in increasing interest for human motion tracking and animation.

5 Conclusions and Perspectives

In this paper we have presented an overview of recent research efforts towards the implementation of virtual avatars and agents, graphically embodied in a 2D and/or 3D interactive worlds, and able to interact intelligently with the environment, other avatars, and particularly with human users.

The field is largely interdisciplinary and complex, involving expertise in computer graphics, animation, artificial intelligence, natural language processing, cognitive and psychological modeling of human-human and human-machine interaction, linguistics, communication, and artificial life, to name a few. Nevertheless, it provides access to a graded series of tasks for measuring (through realistic, even though virtual, evaluating alternatives) the amount of adjustment (as well as intelligence, and achievement) needed for introducing new concepts in the information and communication technology domain. Adaptive, socially enabled and human centred automatic systems will serve remote applications in medicine, learning, care, rehabilitation, and for the accessibility to work, employments, and information.

In the light of the above perspectives, we will expect that these applications will have a beneficial effect on people's lives and will contribute to facilitate communication within dispersed families and social groups.

Acknowledgments

The authors would like to acknowledge the contributions of Aurélie Barbier, Jean Bernard, David Gomez, Christophe Guilmart and Aude Mamouna Guyot-Mbodji to several aspects of the work reported in this paper.

This work has been partially supported by the COST 2102 action "Cross Modal Analysis of Verbal and Nonverbal Communication (www.cost2102.eu)" and the ANR-MyBlog3D project (https://picoforge.int-evry.fr/cgi-bin/twiki /view/Myblog3d/Web/).

References

1. B. Abboud, H. Bredin, G. Aversano, and G. Chollet. Audio-visual identity verification: an introductory overview. In Y. Stylianou, editor, *Progress in Non-Linear Speech Processing*, pages 118–134. Springer Verlag, 2007.
2. M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara. Voice conversion through vector quantization. In *International Conference on Acoustics, Speech, and Signal Processing*, 1988.
3. A. Agarwal and B. Triggs. Recovering 3D human pose from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 44–58, 2006.
4. J. Ahlberg. Candide-3, an updated parameterized face. Technical report, Linköping University, Sweden, 2001.
5. J. Ahlberg. Real-time facial feature tracking using an active model with fast image warping. In *International Workshop on Very Low Bitrates Video*, 2001.
6. I. Albrecht, M. Schroeder, J. Haber, and H.-P. Seidel. Mixed feelings – expression of non-basic emotions in a muscle-based talking head. *Virtual Reality (Special Issue "Language, Speech and Gesture for VR")*, August 2005.
7. L.M. Arslan. Speaker transformation algorithm using segmental codebooks (STASC). *Speech Communication*, 1999.
8. F. Beau. *Culture d'Univers - Jeux en réseau, mondes virtuels, le nouvel âge de la société numérique*. Limoges, 2007.
9. J. Benesty, M. Sondhi, and Y. Huang, editors. *Springer Handbook of Speech Processing*. Springer Berlin Heidelberg, 2008.
10. T. Duy Bui. *Creating Emotions And Facial Expressions For Embodied Agents*. PhD thesis, University of Twente, Department of Computer Science, Enschede, 2004.
11. J. Cassell, J. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjálmsson, and H. Yan. Embodiment in conversational interfaces: Rea. In *CHI'99*, pages 520–527, Pittsburgh, PA, 1999.
12. J. Cassell, S. Kopp, P. Tepper, and F. Kim a nd K. Striegnitz. *Trading Spaces: How Humans and Humanoids use Speech and Gesture to Give Directions*. New york: John wiley & sons edition, 2007.
13. J. Cassell, H. Vilhjálmsson, and T. Bickmore. BEAT : the Behavior Expression Animation Toolkit. In *Computer Graphics Proceedings, Annual Conference Series*. ACM SIGGRAPH, 2001.
14. A. Cheyer and D. Martin. The open agent architecture. *Journal of Autonomous Agents and Multi-Agent Systems*, pages 143–148, March 2001.

15. D. Chi, M. Costa, L. Zhao, and N. Badler. The emote model for effort and shape. In *International Conference on Computer Graphics and Interactive Techniques, SIGGRAPH*, pages 173–182, 2000.
16. G. Chollet, R. Landais, H. Bredin, T. Hueber, C. Mokbel, P. Perrot, and L. Zouari. Some experiments in audio-visual speech processing, in non-linear speech processing. In M. Chetnaoui, editor, *Progress in Non-Linear Speech Processing*. Springer Verlag, 2007.
17. G. Chollet and D. Petrovska-Delacrétaz. Searching through a speech memory for efficient coding, recognition and synthesis. Franz Steiner Verlag, Stuttgart, 2002.
18. T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 681–685, 2001.
19. F. Dornaika and J. Ahlberg. Fast and reliable active appearance model search for 3D face tracking. *IEEE Transactions on Systems, Man, and Cybernetics*, pages 1838–1853, 2004.
20. T. Dutoit. Corpus-based speech synthesis. In Benesty Jacob, Sondhi M. Mohan, and Huang Yiteng (Arden), editors, *Springer Handbook of Speech Processing*, pages 437–453. Springer, 2008.
21. P. Ekman, J. Campos, R.J. Davidson, and F. De Waals. Emotions inside out. In *New York: Annals of the New York Academy of Sciences*, volume 1000. 2003.
22. A. Esposito. Children's organization of discourse structure through pausing means. In M. Faundez et al., editor, *Nonlinear Analyses and Algorithms for Speech Processing, LCNS 3817*, pages 108–115. Springer-Verlag, 2006.
23. A. Esposito. The amount of information on emotional states conveyed by the verbal and nonverbal channels: Some perceptual data. In Y. Stylianou et al., editor, *Progress in Nonlinear Speech Processing, LNCS*, pages 245–264. Springer-Verlag, 2007.
24. A. Esposito and N.G. Bourbakis. The role of timing in speech perception and speech production processes and its effects on language impaired individuals. In *6th International IEEE Symposium on BioInformatics and BioEngineering*, pages 348–356, 2006.
25. A. Esposito and M. Marinaro. What pauses can tell us about speech and gesture partnership. In A. Esposito, M. Bratanic, E. Keller, and M. Marinaro, editors, *Fundamentals of Verbal and Nonverbal Communication and the Biometrical Issue*, pages 45–57. NATO Publishing Series, IOS press, 2007.
26. J.L. Gauvain and L. Lamel. Large - Vocabulary Continuous Speech Recognition: Advances and Applications. In *proceedings of the IEEE*, volume 88, pages 1181–1200, 2000.
27. D. Genoud and G. Chollet. Voice transformations: Some tools for the imposture of speaker verification systems. In A. Braun, editor, *Advances in Phonetics*. Franz Steiner Verlag, 1999.
28. A. Gentes. Second life, une mise en jeu des médias. In A. de Cayeux and C. Guibert, editors, *Second life, un monde possible*. Les petits matins, 2007.
29. R. Gutierrez-Osuna, P. Kakumanu, A. Esposito, O.N. Garcia, A. Bojorquez, J. Castello, and I. Rudomin. Speech-driven facial animation with realistic dynamics. *IEEE Transactions on Multimedia*, pages 33–42, 2005.
30. A. El Hannani, D. Petrovska-Delacrétaz, B. Fauve, A. Mayoue, J. Mason, J.-F. Bonastre, and G. Chollet. Text-independent speaker verification. In D. Petrovska-Delacrétaz, G. Chollet, and B. Dorizzi, editors, *Guide to Biometric Reference Systems and Performance Evaluation*. Springer-Verlag, London, 2008.

31. B. Hartmann, M. Mancini, and C. Pelachaud. Implementing expressive gesture synthesis for embodied conversational agents. In *Gesture Workshop, LNAI*. Springer-Verlag, 2005.
32. D. Heylen, M. Ghijsen, A. Nijholt, and R. op den Akker. Facial signs of affect during tutoring sessions. In J. Tao, T. Tan, and R.W. Picard, editors, *Lecture Notes in Computer Science 3784*. Springer-Verlag, 2005.
33. P. Horain and M. Bomb. 3D model based gesture acquisition using a single camera. In *IEEE Workshop on Applications of Computer Vision*, pages 158–162, 2002.
34. P. Horain, J. Marques-Soares, P. K. Rai, and A. Bideau. Virtually enhancing the perception of user actions. In *15th International Conference on Artificial Reality and Telexistence ICAT05*, pages 245–246, 2005.
35. IV2: Identification par l'Iris et le Visage via la Vidéo. http://iv2.ibisc.fr/pageweb-iv2.html.
36. F. Jelinek. Continuous Speech Recognition by Statistical Methods. In *Proceedings of the IEEE*, volume 64, pages 532 – 556, 1976.
37. A. Kain. *High Resolution Voice Transformation*. PhD thesis, Oregon Health and Science University, Portland, USA, october 2001.
38. A. Kain and M. Macon. Spectral voice conversion for text to speech synthesis. In *International Conference on Acoustics, Speech, and Signal Processing*, New-York, 1998.
39. A. Kain and M.W. Macon. Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction. In *International Conference on Acoustics, Speech, and Signal Processing*, 2001.
40. P. Kakumanu, A. Esposito, R. Gutierrez-Osuna, and O.N. Garcia. Comparing different acoustic data-encoding for speech driven facial animation. *Speech Communication*, pages 598–615, 2006.
41. S. Karungaru, M. Fukumi, and N. Akamatsu. Automatic human faces morphing using genetic algorithms based control points selection. *International Journal of Innovative Computing, Information and Control*, 3(2):1–6, april 2007.
42. A. Kendon. Gesture: Visible action as utterance. Cambridge Press, 2004.
43. M. Kipp, M. Neff, K.H. Kipp, and I. Albrecht. Toward natural gesture synthesis: Evaluating gesture units in a data-driven approach. In *7th International Conference on Intelligent Virtual Agents*, pages 15–28. LNAI 4722, Springer, 2007.
44. S. Kopp, B. Jung, N. Lessmann, and I. Wachsmuth. Max - a multimodal assistant in virtual reality construction. *KI Kunstliche Intelligenz*, 2003.
45. S. Kopp and I. Wachsmuth. Synthesizing multimodal utterances for conversational agents. *The Journal Computer Animation and Virtual Worlds*, 15(1):39–52, 2004.
46. C. Laird, Webster's New World Dictionary, and Thesaurus. *Webster dictionary*. Macmillan, 1996.
47. Y. Li and Y. Wen. A study on face morphing algorithms. http://scien.stanford.edu/class/ee368/projects2000/project17.
48. S. Lidell. *American Sign Language Syntax*. Approaches to semiotics. The Hague ; New York : Mouton, 1980.
49. Samaras D. Metaxas D. Lu S., Huang G. Model-based integration of visual cues for hand tracking. *IEEE workshop on Motion and Video Computing*, 2002.
50. M. Mancini and C. Pelachaud. Distinctiveness in multimodal behaviors. In *Seventh International Joint Conference on Autonomous Agents and Multi-Agent Systems, AAMAS'08*, Estoril Portugal, May 2008.
51. I. Matthews and S. Baker. Active appearance models revisited. *International Journal of Computer Vision*, pages 135–164, 2004.

52. D. McNeill. Gesture and though. University of Chicago Press, 2005.
53. T. Moeslund, A. Hilton, and V. Kruger. A survey of advances in vision-based human motion capture and analysis. *Computer vision and image understanding*, 4:90–126, 2006.
54. K. Moon and V.I. Pavlovic. Impact of dynamics on subspace embedding and tracking of sequences. In *Conference on Computer Vision and Pattern Recognition*, pages 198–205, New York, 2006.
55. R. Niewiadomski and C. Pelachaud. Model of facial expressions management for an embodied conversational agent. In *2nd International Conference on Affective Computing and Intelligent Interaction ACII*, Lisbon, September 2007.
56. M. Ochs, R. Niewiadomski C. Pelachaud C., and D. Sadek. Intelligent expressions of emotions. In *1st International Conference on Affective Computing and Intelligent Interaction ACII*, China, October 2005.
57. M. Padmanabhan and M. Picheny. Large Vocabulary Speech Recognition Algorithms. In *Computer Magazine*, volume 35, 2002.
58. I.S. Pandzic and R. Forcheimer (Eds). *MPEG4 Facial Animation - The standard, implementations and applications*. John Wiley & Sons, 2002.
59. I.K. Park, H. Zhang, and V. Vezhnevets. Image based 3D face modelling system. *EURASIP Journal on Applied Signal Processing*, pages 2072–2090, January 2005.
60. C. Pelachaud, J-C. Martin, E.André, G. Chollet, K. Karpouzis, and D. Pelé. Intelligent virtual agents. In *7th International Working Conference, IVA-2007*, 2007.
61. P. Perrot, G. Aversano, and G. Chollet. Voice disguise and automatic detection, review and program. In Y. Stylianou, editor, *Progress in Non-Linear Speech Processing*. Springer Verlag, 2007.
62. P. Perrot, G. Aversano G., R. Blouet, M. Charbit, and G. Chollet. Voice forgery using alisp: Indexation in a client memory. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 17– 20, Philadelphia, 2005.
63. D. Petrovska-Delacrétaz, A. El Hannani, and G. Chollet. Automatic speaker verification, state of the art and current issues. In Y. Stylianou, editor, *Progress in Non-Linear Speech Processing*. Springer Verlag, 2007.
64. D. Petrovska-Delacrétaz, S. Lelandais, J. Colineau, L. Chen, B. Dorizzi, E. Krichen, M.A. Mellakh, A. Chaari, S. Guerfi, J. D'Hose, M. Ardabilian, and B. Ben Amor. The iv^2 multimodal (2D, 3D, stereoscopic face, talking face and iris) biometric database, and the iv^2 2007 evaluation campaign. In *In the proceedings of the IEEE Second International Conference on Biometrics: Theory, Applications (BTAS)*, Washington DC, USA, September 2008.
65. R. Poppe. Vision-based human motion analysis: an overview. *Computer vision and image understanding*, 108:4–18, 2007.
66. S. Romdhani, V. Blanz, C. Basso, and T. Vetter. Morphable models of faces. In S. Li and A. Jain, editors, *Handbook of Face Recognition*, pages 217–245. Springer, 2005.
67. Z. Ruttkay, H. Noot, and P. ten Hagen. Emotion disc and emotion squares: tools to explore the facial expression space. *Computer Graphics Forum*, pages 49–53, 2003.
68. C. Sminchisescu. 3D Human Motion Analysis in Monocular Video, Techniques and Challenges. *AVSS '06: Proceedings of the IEEE International Conference on Video and Signal Based Surveillance*, page 76, 2006.
69. C. Sminchisescu and B. Triggs. Estimating articulated human motion with covariance scaled sampling. *International Journal of Robotic Research*, pages 371–392, 2003.

70. J. Marques Soares, P. Horain, A. Bideau, and M.H. Nguyen. Acquisition 3D du geste par vision monoscopique en temps réel et téléprésence. In *Acquisition du geste humain par vision artificielle et applications*, pages 23–27, 2004.
71. D. Sündermann and H. Ney. VTLN-Based Cross-Language Voice Conversion. In *IEEE workshop on Automatic Speech Recognition and Understanding*, pages 676–681, Virgin Islands, 2003.
72. D. Terzopoulos and K. Waters. Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 569–579, 1993.
73. M. Thiebaux, A. Marshall, S. Marsella, and M. Kallmann. SmartBody: Behavior realization for embodied conversational agents. In *Seventh International Joint Conference on Autonomous Agents and Multi-Agent Systems, AAMAS'08*, Portugal, May 2008.
74. K.R. Thorisson, T. List, C. Pennock, and J. DiPirro. Whiteboards: Scheduling blackboards for semantic routing of messages and streams. In *AAAI-05 Workshop on Modular Construction of Human-Like Intelligence*, Pittsburgh, Pennsylvania, July 10 2005.
75. D. Traum. Talking to virtual humans: Dialogue models and methodologies for embodied conversational agents. In I. Wachsmuth and G. Knoblich, editors, *Modeling Communication with Robots and Virtual Humans*, pages 296–309. John Wiley & Sons, 2008.
76. N. Tsapatsoulis, A. Raouzaiou, S. Kollias, R. Cowie, and E. Douglas-Cowie. Emotion recognition and synthesis based on MPEG-4 FAPs in MPEG-4 facial animation. In Igor S. Pandzic and Robert Forcheimer, editors, *MPEG4 Facial Animation - The standard, implementations and applications*. John Wiley & Sons, 2002.
77. E. Turajlic, D. Rentzos, S. Vaseghi, , and C.-H. Ho. Evaluation of methods for parametric formant transformation in voice conversion. In *International Conference on Acoustics, Speech, and Signal Processing*, 2003.
78. S. Turkle. *Life on the screen, Identity in the age of the internet*. New York: Simon and Schuster, 1997.
79. R. Urtasun, D.J. Fleet, and P. Fua. 3D people tracking with gaussian process dynamical models. In *Conference on Computer Vision and Pattern Recognition*, pages 238–245, New York, 2006.
80. H. Vilhjalmsson, N. Cantelmo, J. Cassell, N. E. Chafai, M. Kipp, S. Kopp, M. Mancini, S. Marsella, A. N. Marshall, C. Pelachaud, Z. Ruttkay, K. R. Thórisson, H. van Welbergen, and R. van der Werf. The behavior markup language: Recent developments and challenges. In *7th International Conference on Intelligent Virtual Agents, IVA'07*, Paris, September 2007.
81. L. Wiskott, J. M. Fellous, N. Krüger, and C. von der Malsburg. Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 775–779, 1997.
82. G. Wolberg. Recent advances in image morphing. In *Computer Graphics Internat*, pages 64–71, 1996.
83. J. Xiao, S. Baker, I. Matthews, and T. Kanade. Real-time combined 2D+3D active appearance models. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 25–35, 2004.
84. H. Ye and S. Young. Perceptually weighted linear transformation for voice conversion. In *Eurospeech*, 2003.
85. B. Yegnanarayana, K.Sharat Reddy, and S.P. Kishore. Source and system features for speaker recognition using AANN models. In *International Conference on Acoustics, Speech, and Signal Processing*, 2001.

86. S. Young. Statistical Modelling in Continuous Speech Recognition. In *Proceedings of the 17th International Conference on Uncertainty in Artificial Intelligence*, Seattle, WA, August 2001.
87. V. Zanella and O. Fuentes. *An Approach to Automatic Morphing of Face Images in Frontal View in MICAI 2004: Advances in Artificial Intelligence*. Springer Berlin - Heidelberg, 2004.