TTS-driven Expressive Embodied Conversation Agent EVA for UMB-SmartTV

Matej Rojc, Marko Presker, Zdravko Kačič, and Izidor Mlakar

***Abstract*—** The main goal of using non-verbal modalities together with the general text-to-speech (TTS) system is to better emulate human-like course of the interaction between users and the UMB-SmartTV platform. Namely, when human-TV interaction is supported by TTS only, the interactions tend to be still less functional and less human-like. In order to achieve more advanced interaction, and more human-human like, the virtual agent technology as a feedback interface has to be introduced. In this way more appropriate social responses from the UMB-SmartTV through personification of the TTS system, named PLATTOS, can be produced and close to human-human-like communicative behavior may be invoked. Verbal and co-verbal gestures are linked through complex mental processes. Understanding of attitude, emotion, together with how gestures (facial and hand) and body movements complement, or in some cases, override any verbal information produced by the TTS system, provides crucial information for modeling both the interaction and the embodied conversational agent's (ECA) socially-oriented responses. The social responses of the TTS system fused with ECA can then be presented to the user in a more human-like form, using not just audio but also facial expressions, such as: facial emotions, visual animation of synthesized speech, and synchronized head, hand, and body movements. In the paper a novel TTS-driven behavior generation system is proposed to be used for IPTV platforms. The behavior generation engine is implemented as a service and used by UMB-SmartTV in a service-oriented fashion. The behavior generation engine fuses both, speech and gesture production models, by using FSMs and HRG structures. Selecting the shape and alignment of co-verbal movement for embodied conversational avatar, named EVA, are based on several linguistic features (automatically extracted from the input text), and several prosodic features (symbolic and acoustic features produced within the TTS engine). Finally, the generated speech and co-verbal behavior are animated by embodied conversational agent's engine and represented to the user within the UMB-SmarTV user interface. In this way, personificated TTS system PLATTOS, integrated within the UMB-SmartTV system, enable more advanced, personalized, and more natural multimodal-output-based human-machine interface.

R. M. Author is with the Faculty of Electrical Engineering and Computer Sicence, University of Maribor, Maribor 2000 Slovenia (e-mail: matej.rojc@uni-mb.si).

P. M. Author is with the Faculty of Electrical Engineering and Computer Sicence, University of Maribor, Maribor 2000 Slovenia (e-mail: marko.presker@uni-mb.si).

K. Z. Author is with the Faculty of Electrical Engineering and Computer Sicence, University of Maribor, Maribor 2000 Slovenia (e-mail: kacic@uni-mb.si).

M. I. Author is with the Panevropa d.o.o., Maribor 2000 Slovenia (e-mail: izidor@panevropa.com).

***Keywords*—**embodied conversational agents, Smart TV, UMB-SmartTV, behavior generation, human-machine interfaces, EVA-framework.

I. Introduction

There exist many possibilities for development on smart TV platforms and many interactive services can already be provided. These platforms (e.g. [1, 2, 3]) contain more and more sophisticated applications and features, accessible through more and more complex menu structures. The human-TV interaction mechanisms should, therefore, respond in terms of efficiency and higher degree of naturalness. The platforms should also employ different interaction techniques ranging from tactile to audio/visual interaction. More natural and more personalized TV units can be implemented by virtual agent technology, incorporating affective and intelligent talking avatars, responding in a human-like fashion [4, 5].

In developing advanced and more natural output for smart TV platforms, text-to-speech synthesis (TTS) systems can be used for synthesizing speech for any general text originating from IPTV services (e.g. EPG, VOD, broadcasts news, etc.), or other web services and system messages [6]. Nevertheless, audio channel itself is not sufficient anymore. Embodied conversational agent paradigm is being effectively integrated into different user interfaces (UI), including Smart TV platforms [7, 8]. Namely, ECAs may further personificate Smart TV platforms [9, 10] and may have a huge impact on interactivity and personalization of IPTV systems and other services provided by Smart TV platforms.

The integration of non-verbal modalities together with text-to-speech systems also better emulate the natural course of the dialogue between users and the IPTV system. Such integration enables users to interact with a virtual person, and makes them feel more comfortable when "talking" to a Smart TV UI. The next reason is hidden in issues that generally occur whilst using human-machine interaction systems: repetition and misinterpretation of speaking terms are common features in human-machine interaction (HCI). These features usually lead towards less-functional and less-efficient interaction that degrades user experience [11]. The misinterpretation and repetition are also a commonality in current smart TV units (e.g. Samsung, Apple). When more natural social responses, by using embodied conversational agents (ECA), are available, users tend to more readily respond with emotive socially-colored responses. In this way human-human-like

communicative behavior may be evoked, giving the spoken dialogue system within smart TV platforms the ability to shape and adjust the dialogue to its own rules.

Therefore, the TTS systems and believable conversational agents (ECAs) can be used together to achieve a higher level of social interactivity [10]. Understanding of attitude, emotion, together with how gestures (facial and hand) and body movements complement, or in some cases, override any verbal information increases the viability of social responses. The response of a service may, therefore, be represented to the user in a form of synchronized speech, facial expressions, and movements of head, hands, and body. In this way a personificated TTS system is formed. Such a system enables the development of more advanced and personalized IPTV and interactive services that can be used for today's smart TV units.

In this paper an integration of TTS-driven embodied conversational agent technology into Smart TV platforms is presented. Although ECAs and synthetic "communicative" behavior have already been researched for some time, the co-alignment of speech and non-verbal expressions still represents an important and challenging research problem. Namely, the correlation between verbal and non-verbal signals in communication (co-verbal expressions), and the process involved in co-speech gesture production, originates in semantic, pragmatic, and temporal synchronization of the multimodal-content [12-13]. Some co-verbal gestures, such as: iconic expressions [14-15], symbolic expressions [16], and mimicry [17] are tightly interlinked with speech. These gestures may be identified by linguistic (semantic) properties of the general input text, e.g. by considering word-type, word-type-order, word-affiliation, etc. But many co-verbal gestures especially those representing communicative functions (e.g. indexical and adaptive expressions [18]), have less (if at all) evident semantic or linguistic alignment with the text. Nevertheless, they may still be identified by linguistic fillers [19], turn-taking, and directional signals.

II. UMB-SMARTTV

In Figure 1 is presented the architecture of the UMB-SmartTV system that is based on IMS infrastructure. The system consists of STB, TV server, VOD server, IMS core, presence and XCAP server (Kamailio [20, 21]), environment controller gateway (raspberry pi), and distributed DATA system [22], used for multimodal platform within the UMB-SmartTV system. The UMB-SmartTV system can be operated by using standard input devices (keyboard, mouse, remote controller etc.), by using mobile devices (e.g. mobile phone, tablet), or by speech. VoD server, IMS core, XCAP server, presence server, and TV server are running on Linux-based operating system (Ubuntu). STB software runs on Windows 7, XP or Linux, and distributed DATA system's servers on Windows XP. UMB-SmartTV system consists, therefore, of several hardware/software blocks, but unifying them into a powerful multimodal media platform. The Content core represents an application server and takes care for content production and content presentation, like Live TV, VOD, RSS etc. The client-sided service and GUI (the IMS client) are then implemented by XBMC (Xbox Media center) [34]. The next block represent IMS core that represents multimedia service platform architecture. Additionally, IMS core implements standard IMS functionalities, such as: user registration, subscription and management, session management, triggering, routing, interaction with NGN services, and QoS control. Multimedia services are served by distributed system for providing automatic speech recognition and text-to-speech synthesis (module servers), virtual assistants (ECA server), home automation, and personalization. Finally, environment controller implements means to control several devices in the environment/household using raspberry pi platform and Z-wave mesh networking technology. In the next section, multimodal platform integrating virtual agent technology is presented into detail.

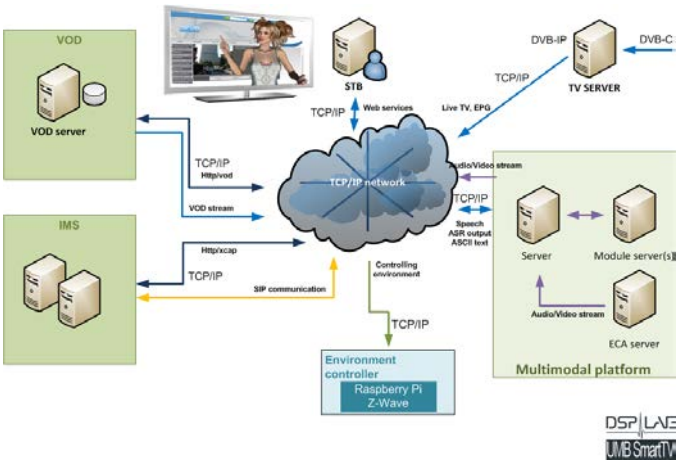

Fig. 1 architecture of the UMB-SmartTV system

III. MULTIMODAL PLATFORM

There can be several users communicating with the Smart TV unit. Those users expect in general user's specific behavior of Smart TV unit when interacting with it. Therefore, IPTV systems should be operated with distributed multimodal platforms that are able to implement and perform user's specific behavior for several users. Further, distributed system architectures have to be able to process events, triggered by users, constantly and usually in asynchronous time instants. After events are detected and processed, those tasks have to be performed, as defined/allowed by system behavior scenario for specific user. Multimodal platform based on distributed DATA system fulfills all these features. It consists of main DATA server (managing unit), and several DATA module servers, used for running several engines, such as: text-to-speech synthesis engine (TTS), speech recognition engine (ASR), spoken dialogue engine, and embodied conversational agent's animation engine. And all DATA modules are event-based finite-state machines, based on Java programming language. The dialog manager engine drives the interaction between users and

UMB-SmartTV system. Depending on recognized words, current state, and pre-defined user-specific scenario, it sends system's messages and general text to the TTS engine. TTS engine then generates corresponding audio, and also behavior script for animating ECA EVA. Both outputs are then send via DATA server to the ECA server, where animation engine, called Panda, is running as TCP/IP server. After receiving both needed outputs, ECA server produces audio/video stream that is sent to the XBMC-based interface [23], as seen in Figure 1. In this way the IPTV UMB-SmartTV system is supported by multimodal output, combining TTS and ECA engines' outputs.

In the next section, the proposed TTS-driven conversational behavior generation system for UMB-SmartTV system is presented in more detail.

IV. TTS-DRIVEN CONVERSATIONAL BEHAVIOR GENERATION SYSTEM FOR UMB-SMARTTV

In the UMB-SmartTV system the relationships that link general text and co-verbal gesture are established within common engine, in order to better synchronize verbal and non-verbal behavior in both meaning and time (Figure 2). Additionally, the system synchronizes the co-verbal expressions in a way that the meaningful part of a gesture co-occurs with the most prominent segment of the accompanying generated speech [24]. And all features that are driving the co-verbal behavior, are deduced completely from general (semantically, or otherwise untagged) text. The processing steps for planning and generating non-verbal behavior involve semiotic grammar, gesture dictionary, and lexical affiliation that are included into the behavior generation system as external resources (language dependent). The behavior system at the end transforms the co-verbal expressions into a form understandable to ECA-based behavior realization-engines running on ECA server (supporting mark-up languages, such as: BML [25], and EVA-SCRIPT [26, 27]).

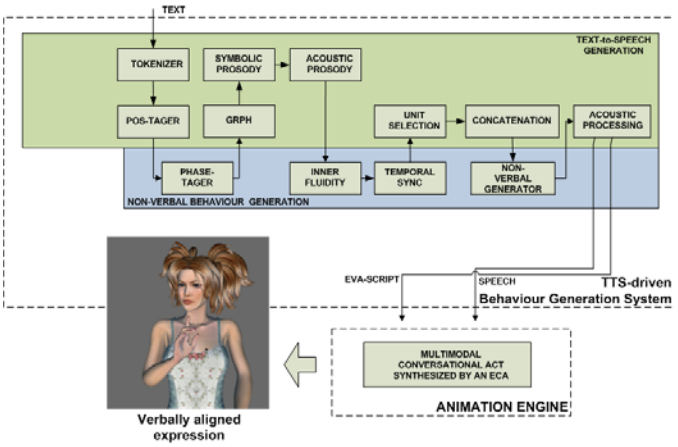

Fig. 2 TTS-driven behavior generation system

The TTS-driven system for generating co-verbal conversational behavior, as presented in Figure 2, is based on core TTS system, named PLATTOS [28, 29]. The system is modular, time and space efficient, and flexible. Further, the language dependent resources are separated from the language independent conversational behavior engine, by using FSM formalism and CART models. Further, well established queuing mechanism allows for flexible, efficient and easy integration of several modules, used for synthesis of non-verbal expressions symbolically and temporarily aligned with speech. In Figure 2, the following modules are additionally added to the core TTS engine: phase-tagger, inner-fluidity, temporal-sync, and non-verbal generator module.

The phase-tagger module is used for the symbolical synchronization of verbal and non-verbal behavior and the inner-fluidity module is used for specifying the inner-fluidity of the conversational expression(s). The temporal-sync module is then used to temporally align the propagation of movement with the generated pronunciation of verbal content. The final non-verbal generator module is used for transforming the generated behavior into procedural behavior description that can be then animated by animation engine on ECA server. In this way, the system is completely TTS-driven, and it benefits from the core TTS system, and from its underlying predicted linguistic and prosody features, as used for generation of speech from general text (e.g. stress, prominence, phrase breaks, segments' duration, pauses, etc.). In this way any information about the form of movement (content) and about the co-alignment with generated speech can be extrapolated from the general text. Within IPTV systems, it is also important that system's outputs in different virtual/physical interfaces are running at interactive speeds. In the presented system, the multimodal output generation is efficient since the TTS-driven behavior system synthesizes the non-verbal behavior description, and corresponding speech signal, simultaneously. In the following subsections each non-verbal deque of the TTS-driven behavior generation system will be presented into detail.

A. Phase tagging deque

This deque is responsible for the symbolical synchronization and it has to identify the so-called semiotic phrases (the words carrying most meaning - the meaningful words), and the shapes that could further illustrate or emphasize the meaning of the indicated meaningful words. For identifying these meaningful words and word-phrases, the semiotic grammar is used. The semiotic grammar is also closely interlinked with an off-line constructed gesture lexicon (gesture affiliates). Previous module (POS-tagger) provides the needed input information. All data extracted from input text are stored in the form of heterogeneous relation graph (HRG) (Figure 3), where they can be easily find and accessed, and transferred between several modules in the system. This module creates within a HRG structure additional ContentUnits relation layer, where generated Content units (CUs) can be stored and then used in the following system's modules. Namely, these units are used to establish the relationships between the shapes manifested within the movement-phases. This deque performs symbolic synchronization (verbal-trigger indication, and content

selection) by performing semiotic tagging, semiotic processing, and matching process. Namely, the deque processes POS-tagged word sequences on the sentence level [28], and matches them against semiotic grammar's rules and relations stored in the off-line constructed gestural dictionary. Here, FSMs are used in order to identify the meaning and to select the physical representation of the meaning, and to also predict how the propagation of meaning could be performed over the specific text sequence (particularly in the preparation movement-phase, and within concepts of repetitive, circular, enumerative gestures). The phase tagging deque, therefore, performs synchronization of the form since it defines what the meaningful phrase is and how it is conveyed through bodily manifestations (e.g. what to convey, and how to convey it [30]).

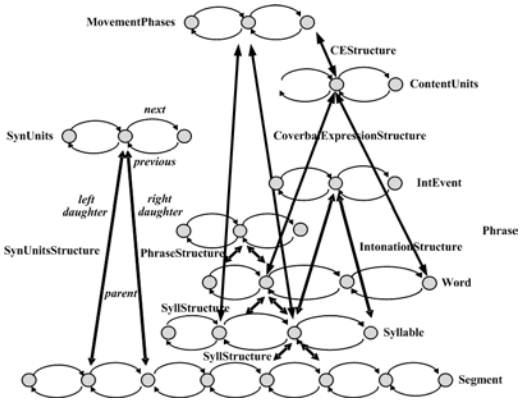

Fig. 3 the HRG structure for storing verbal and non-verbal information within behavior generation system

B. Inner fluidity deque

The inner fluidity deque has to perform the alignment of the propagation of co-verbal expression, and the input-text (in the syllable level). Within HRG structure it creates additional layer, named MovementPhases. This layer is used for describing the relation between propagation of co-verbal movement, and the input text. Firstly, the starting and ending points of CUs (generated by the phase tagging deque) are compared against prosodic word phrases (logical content segments), as predicted by the symbolic prosody deque (module used for text-to-speech synthesis). Namely, these prosodic phrases indicate those text-sequences with a complete meaning [28]. Further, sentences containing more than one prosodic phrase can indicate additional explanation, emphasis, or even negation of the preceding meaning. The deque's process "search for phrase breaks" adjusts (extends, or even removes) the starting and ending points of the CUs, generated before. Basically, it aligns CUs with their predicted prosodic counterparts. A general rule that we applied here is that each prosodic phrase can contain only one (or even none) co-verbal expression. The movement-propagation of each co-verbal expression must also be maintained within the indicated prosodic phrase. The next process in this deque ("searching of emphasized words") identifies the emphasized words, where emphasized word is identified as a word that contains a syllable assigned with a primary accent (PA). This information is predicted by the symbolic prosody deque of the TTS system's module (by using CART models), and assigned to the specific syllable within the prosodic phrase. The third process is then "searching of stressed syllables" that aligns the stroke movement phases according to the syllable having a PA. Namely, the starting and ending points of the stroke phase has to be determined by the beginning of the emphasized word and by the end of the syllable. The fourth process is "align stroke" that has to align the preparation movement-phases. Here, the CUs already contain the definition of the shape preceding the meaningful shape (e.g. the "initial" physical manifestation from which the body transforms throughout the stroke phase). In the fifth process, named "align preparation", we finally align the preparation movement phase. The preparation phase is identified by the first word with an unaccented (NA) syllable preceding specific prosodic gesture trigger. The starting and ending points of the preparation phase are, therefore, determined by the NA syllable, and the end of the "preparation" word. The sixth process performed in this deque is "align hold/retraction" that has to align hold (both pre- and post-stroke) and retraction movement phases. In this case the retraction movement phase is determined by the last meaningful phrase; by the word that contains NA syllable and is preceding the major phrase break level tagged as B3, or is preceding a longer pause.

The MovementPhases layer in the HRG structure stores the movement structure of the observed sequence and it outlines what shape should manifest over specific words, where should the movement be withheld, and where retracted to its neutral (rest) state. However, this deque still does not specify any temporal boundaries for movement generation, what is actually needed at the end. Additionally, at this level there can be several repeated holds within the given movement sequence which should later either be filtered, or merged. Therefore, the temporal-sync deque has to be included and performed next.

C. Temporal-sync deque

This deque has to temporally align verbal and non-verbal behavior. Now this is possible, since temporal information is already predicted at phoneme/viseme level by the acoustic prosody deque and stored as units in the Segments relation layer in the used HRG structure. Additionally, the Segments relation layer stores temporal information about predicted pauses (inserted as sil units). Generally, this temporal information in the Segments relation layer is used by the core TTS engine. But in the non-verbal behavior generation system, by using temporal information, the duration of each movement phase can also be specified. The deque uses information from the following HRG's layers: ContentUnits relation layer, Segments relation layer, and MovementPhases relation layer. This deque at the end produces new units, named Phase units

(PUs), and are stored in the MovementPhases relation layer of the HRG structure. It also adds additional attributes to the CUs, stored within ContentUnits relation layer.

The first process is "filter process" that searches for those sil units that have predicted durations 0 ms (can be sil units, and/or phase-breaks). Namely, each sil unit can represent a hold movement-phase in the movement structure (MovementPhases relation layer). In order to filter out "false" hold-movement phases, the vertical HRG's relations between sil units and the hold movement-phases, are deleted, and the corresponding Phase units removed from the MovementPhases relation layer. This process also takes care for merging repeated hold-movement phases into a single hold. The starting and ending point of the merged hold are in these cases determined by the starting point of the first hold and by the ending point of the last hold. The second process is "Align CE" that has to temporally align each conversational expression (CE) with the input text. The temporal values determined for the CUs and PUs are calculated from the temporal values of their children (phoneme, and sil units), segment units stored in the Segment layer of the HRG structure. There are two types of units at the output of this deque that store all the information necessary for the recreation of the generated co-verbal behavior plan by using an embodied conversational agent. The CUs contain global temporal structure for the corresponding conversational expressions, whereas the PUs contain local information regarding overlaid shapes.

In the following subsection the last non-verbal deque is presented. This deque has to transform CUs and corresponding PUs into procedural descriptions of synthetic behavior (behavior-plan), as required for virtual recreation of non-verbal behavior (including hand/arm gestures, and lip-sync) by using synthetic embodied conversational agents on ECA server.

D. Non-verbal generator deque

Most ECA-based animation engines recreate non-verbal behavior based on some procedural animation description mark-up language, such as: BML [25], and EVA-SCRIPT [26, 27]. All mark-up languages require at least temporal specification (e.g. relative position, duration etc.) of behavior, and the description of shape (provided in at least abstract notation). These behavior (animation) descriptions have then to be fed to the animation engine, and recreated by a synthetic ECA (performed on ECA server).

The non-verbal generator deque transforms the information, as generated and stored in the common HRG structure, into a form understandable to the animation engine. Namely, this deque has to transform the HRG structure into EVA-SCRIPT-based behavior descriptions, and is supporting both lip-sync, and co-verbal gesture animation processes. Nevertheless, since the HRG structure stores very detailed information on non-verbal behavior, also the transformation into other mark-up languages is possible and straightforward. This deque uses as input the CUs that are stored in the ContentUnits relation layer, and the PUs that are stored in the MovementPhases relation layer in the HRG structure. And the output is then a behavioral script, written in EVA-SCRIPT animation description mark-up language. The EVA-SCRIPT-based descriptions contain hierarchical structures that can very precisely describe the configuration of the movement controllers, and the duration in which this configuration is to be reached. The implemented mechanisms are based on forward kinematics, and frame-based key-pose specification. Furthermore, in order to re-create (animate) the conversational expressions as described and stored in the HRG structure, those shape models, determined by the PUs attributes (e.g. rhshape, rashape, lhshape, and lashape), are selected from the external gestural dictionary. These models already contain the movement controllers' configurations, and must only be temporally adjusted according to the temporal specification, as specified in the CUs and PUs. In this case the shape models can be directly accessed by the animation engine, and do not need to be further specified in more detail.

As can be seen in Figure 4, the non-verbal generator deque transforms the symbolically and temporally aligned non-verbal behavior into procedural animation (EVA-Script behavior event). The automatic process traverses through generated CUs, recalls the aligned PUs, and finally generates the XML description. The shapes specified in these XML descriptions are then used (Figure 4, right-hand-side) to recreate the selected shapes on embodied conversational agent ECA (e.g. EVA [31]), and used as multimodal output in the UMB-SmartTV system (audio/video stream served by ECA server).

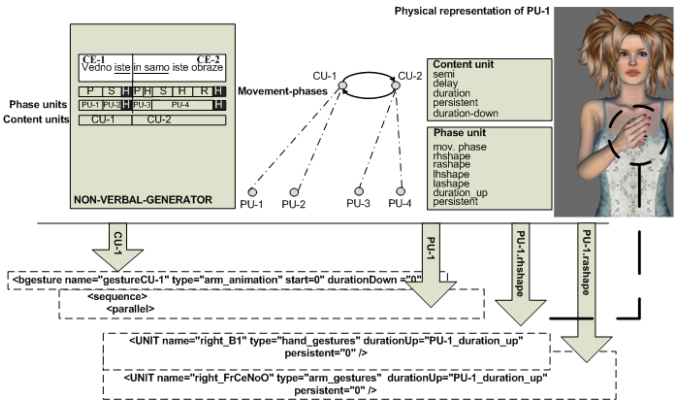

Fig. 4 transformation of data in the HRG structure into procedural animation specification

The proprietary EVA-framework is used for animating ECA, and has been developed to evoke a social response in human-machine interaction. It is a python-based software environment that can convert generated TTS-driven system's output into audio-synchronized animated sequences. ECA's provided by EVA framework can generate social responses in the form of facial expressions, gaze, head and hand movement and, most importantly, in the visual form of synthesized speech. The EVA framework provides a description script, an

animation engine and articulated 3D models, and provides visual representation of synthesized speech sequences in the form of different types of video streams (in addition to synthesis into a video file/screen) [36].

In the next section the animation framework needed for the UMB-SmartTV system is presented into more detail.

V. EVA-Framework

An understanding of attitude, emotion, together with how gestures (facial and hand) and body movement complements, or in some cases even overriding verbal information, provides important information about modeling interactive management, when generating natural human-machine interaction. Personification of the TTS system PLATTOS, not only relates to the transformation of a TTS system's output into ECA's visually-presented articulation within the mouth region (visualizing verbal behavior), but also to the visualization of non-verbal behavior. Therefore, the most natural way to visualize and emulate face-to-face conversation, both for verbal and non-verbal information is to translate it into a human-body representation. ECA implementations can be in the form of e.g. talking heads [37], to agents that can move and use the whole representation of the human body like in [40]. Many implementations of ECA's can in one way or another emulate natural human behavior and evoke emotional and social responses [39], describe the generation of emotional responses and the recognition of emotions by humans, and the additional adaption of ECA's personality to that of the human. E.g. Poggi and Pelachaud generate communicative behaviors on the basis of speech acts and concentrate on one facial expression and speech act performatives, which are a key part of the communicative intent of a speaker, along with propositional and interactional acts [38]. And any conversational action in any modality can result in several (sometimes contradictive) communicative goals. The general architecture of a system that can be used to visualize and to personificate a general TTS system is formed as shown in Figure 5.

Fig. 5 architecture of a general ECA visualization system

Different input modalities are combined into different behavioral events. And different input modalities are commonly generated as abstract behavior descriptions provided in XML based description schemes such as Affective Presentation Mark-up Language (APML) [41], and behavioral mark-up language (BML) [42]. These description languages are used to describe any movement/action realized within the scope of human-machine interaction. The input manager, commonly referred to as a behavior modeler, processes different modality-dependent inputs, and transform them into a time-referenced set of behavioral events. Nevertheless, a time-scheduling process has to synchronize verbal with non-verbal behavior, such as facial expressions, head movements, gaze and head gestures. The behavioral events (or behavior controllers) then form motion's speech-synchronized descriptions that have to be transformed into movement of the ECA's articulated model. And different types of articulated bodies are used (Güdükbay et al., 2008). Namely, 3D models can be grouped into:

- Stick figure models: models based on sets of rigid elements, connected to joint chains.
- Surface models (mesh-based models): an upgrade of stick figure models, but in this case, a polygonal mesh-layer (skin) is applied on the skeleton chains.
- Volumetric models: these models use simple volumetric primitives, such as: spheres, cylinders and ellipsoids, in order to construct the body shape.
- Multilayered models (muscle-based models): these models present anatomically-correct models. Nevertheless, the animator of such models introduces different kinds of constraints to the relationship between layers.

The ECA realization engine in Figure 5, is used to store the articulated models of different ECA's bodies, and to apply behavioral events in the form of different transformations on the control units (those parts of the articulated model used to generate movement). These transformations then result in animated movement. The animation technique's type depends on the type of articulated model. In general, such animations are performed in the form of skeletal joint transformations and morphed-shaped transformations.

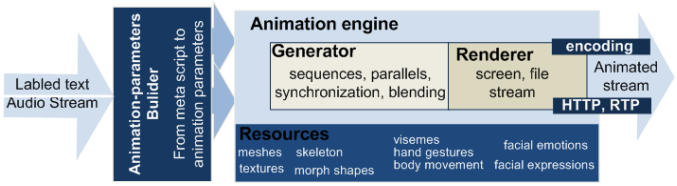

Fig. 6 EVA-framework architecture

The proprietary EVA-framework [36], has been developed in order to evoke a social response in human-machine interaction, and is a python-based software environment that converts a PLATTOS TTS system's output into audio-synchronized animated sequences. Further, ECA EVA provided by EVA-framework can generate social responses in the form of: facial expressions, gaze, head, and hand movement and, in the visual form of synthesized speech. The EVA-framework consists of animation engine and articulated 3D models, and provides visual representation of general synthesized speech sequences in the form of different types of

video streams (in addition to synthesis into a video file/screen). Figure 6 outlines the modular architecture of the EVA framework.

As presented in the non-verbal generator deque, the TTS-driven behavior generation system produces TTS output according to the EVA-framework's specifications, based on the EVA-Script XML scheme as proposed in [36]. These schemes specify the desired ECA facial animations, and body movements. The animation engine has to translate them onto the articulated 3D model EVA in the form of animated movement. A TTS system's behavior generation output contains acoustic, linguistic, and temporal information about input texts that are realized in a two stage visualization concept. The first stage is named 'Animation building' and the second 'Animation realization'. The first stage transforms TTS output (in the form of the EVA script) into animation parameters that are mapped to different control units of the ECA's 3D articulated model. Then the transformation from abstract to animatable content is performed by using the Animation parameter's builder. This transformation can be described as interfacing several XML tags using different ECA resources.

Each ECA has two types of resources. As seen in Figure 6, the 3D multi-part actor resources contain several 3D-submodels of body, e.g. hair-style, eyes, teeth, dresses, etc. And each 3D-submodel is associated with its corresponding textures, polygonal meshes, and sets of different control units, e.g. morphed shapes, and skeletal chains. The template resources contain several behavioral templates, written in the EVA script, that specify common articulation of an ECA, triggering content words to gesture translations (what is a common gestural sequence when a certain word occurs), and other distinctive ECA's features (e.g. eye-blinks, probability of gesturing, etc.) making specific ECA as an individual 'person'. The Animation parameters builder translates the labeled text by interfacing each EVA script's tag with a control unit, or behavioral template, and forms different groups of movements. And each group of movement is defined by semantic (which control units in which order), temporal (the duration of stroke, hold, and retraction phases), and spatial features (ending position of the control unit). Finally, the Animation realization phase takes care for transforming animation parameters into animated sequences. The animation parameters are those raw data that specify how the Animation engine should move different control units. And the EVA-framework *animation engine* takes care of animating and rendering the obtained animation parameter sets. The engine, based on the Panda 3D game engine [43], actually transforms the animation parameter sets into corresponding sequential and/or parallel movements of control-points (bones, or morphed shapes). And each control-point within 3D space can be moved either by 3D transitional or 3D rotational vectors (specified in MPEG4 standard). It is important that the Forward kinematics and animation engine's generator are able to provide procedures for the synchronization of movements, and the animation-blending technique that has to be used on those animated segments that have to be controlled by different gestures at the same time, e.g. animating smile and viseme simultaneously. In this case most of the influence is given to the viseme animation, and only a small portion is left to the facial gesture smile. The animation groups the control units into sets of sequential and concurring movements, and adjusts each movement with its associated temporal and spatial features.

Further, the EVA-framework presumes that no movement is linear and should be interpolated against its interpolation curve. Therefore, EVA-framework provides three types of non-linear interpolation for each movement: EaseIn (slow-start and ramp-to-full, abrupt finish), EaseOut (starts with full speed and in the last n frames decelerates to a slow stop), and EaseInOut (starts slowly, ramps to full speed and after the constant phase, if it exists, slowly decelerates to full stop).

The rendering process is frame-based. Therefore, at any given frame the animation can be stopped/paused, or re-adjusted to its specified temporal/spatial features. The EVA scripts as produced by TTS-driven behavior generation system, describe both verbal and non-verbal, independently. The verbal behavior is stored within speech XML tags (named speech), and the non-verbal behavior is contained within <fgesture> and <bgesture> tags, describing facial expressions, and different body gestures. The verbal parameters are described by the semantic, temporal, and articulation features of a sequence. And non-verbal behavior parameters involve describing the presence level of facial expressions, and different body gestures. The non-verbal behavior is driven by TTS-driven behavior generation system's output, such as: emphasis, phrase/word breaks, and key phrases (e.g. dialogue discourse markers). And the non-verbal feature allocators, <fgesture> and <bgesture> unify a set of those control units, assigned to control different parts of the body.

Further, the facial expressions contain control units that can be physically assigned to the human face, e.g. control units such as lower-jaw, mouth corners, etc. In similar way, body gestures allocate control units, such as: left elbow, neck, control units for fingers, etc. Additionally, the left and right-eye control units are assigned to the body gesture group. By describing the temporal and spatial features of movement in the form of sequential or parallel groups, the EVA-framework enables hierarchical levels of animation for both <fgesture> and <bgesture> objects. In this way, each movement is built from different control units with either sequential or concurrent movement.

The EVA framework not only enables highly realistic human-machine interaction, but can also evoke emotional and social responses existing in face-to-face human-human interaction. Facial expressions motion-templates and control-keys are defined based on MPEG4-FAP and according to Ekman [44], and each facial expression/emotion is generated by combining different sequential/concurrent transformations of different FAPs. EVA-script section describes a description of 'motion' as a set of temporally-defined end-poses, whereas

in-between the key-frames is calculated and interpolated automatically. Figure 7 presents a description of facial expression-templates within behavioral event. The facial region only describes the fundamental temporal and spatial features of transformation, and the structural relations between different FAP-based control-keys. When EVA-framework processes a description, it transforms different temporally-described sequential/parallel behavioral event-segments into a synchronized, fluid stream of facial motion. To retain the naturalness of speech co-articulation and at the same time to prevent "jerky" expressions, the animation blending-techniques are used. Additionally, animation blending-techniques are used to present different types of complex emotions, such as: emotion masking, mixed expressions, and qualified emotions.

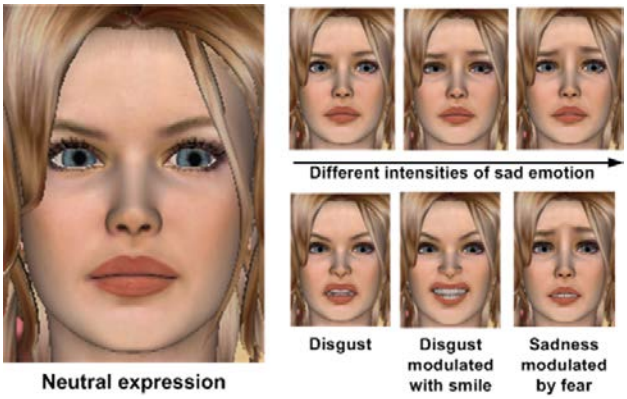

Fig. 7 facial expression modulation using EVA-Script

In [44], it is distinguished between modulating, falsifying, and qualifying an expression. The modulation of an expression is within EVA-framework realized by using methods of intensification, or de-intensification. Namely, when falsifying an expression, a person usually simulates it, or masks it (overlays a fake emotion over truthful emotion). And the qualification of emotion usually adds a fake expression in order to modulate the emotion (e.g. a smile modulates a sad facial emotion). The EVA-framework realizes these facial modulation methods by using animation blending-techniques combined with the modulation of different expressive features. Further, the intensification/ de-intensification and qualification of expression can be realized by power (e.g. stress attribute), and temporal (durationUp, durationDown, persistence attributes) properties. By considering the hierarchical natures of behavioral events, facial-modulations can be also modulated over time and structured to co-inside with different speech-related features, e.g. a sad facial-emotion can be modeled by a smile in respect of certain phrases. A smile can be part of a new "interruptive" behavioral event (independent motion segment), in order to simulate reactive behavior, or pre-planned, and inserted within the hierarchical structure of "planned" behavior. Figure 7 presents different facial expression-modulations that can be realized by following the EVA-Script scheme. The left-hand-side of Figure 7 presents a neutral expression, and the right-hand-side presents different modulations of an expression (e.g. intensification of sad emotion, from low-to-high intensity, and qualification of different emotions using different facial expressions/emotions). In this way, the ECA EVA by using EVA-framework can express also complex facial expressions and emotions based on different control-keys and emotion-templates relating to MPEG-4 FAP and Ekman's six base emotions (anger, disgust, fear, joy, sadness, and surprise).

VI. RESULTS

In order to use the presented TTS-driven behavior generation system in the UMB-SmartTV system, and to evaluate the quality and naturalness of the generated output, we have to prepare additional external language dependent resources. We annotated over 35 minutes of the proprietary video corpus, and created the gestural dictionary containing up to 300 distinct conversational configurations of arms (100) and hands (80), already described in form of EVA-SCRIPT shape models. The shape models varied in structure and intensity of the shape. And the shape configurations are (based on the annotation and literature) also linked to the verbal information (words and phrases).

Fig. 8 ECA integrated in the XBMC interface of the UMB-SmartTV system

In the on-line system they can be automatically selected and also temporally adjusted during the presented non-verbal behavior generation process. In preparing external resources, we relied on findings presented in literature (e.g. [32-33]), however, we have also ensured that manually selected meaningful words in the annotation sequences had at least one representative affiliate stored in the gesture dictionary that can be accessed either based on semiotic or implicit rules. In the on-line UMB-SmartTV system (Figure 1), raw text is sent first to the TTS-driven behavior generation system, ran by corresponding DATA module server. When outputs are generated, they are sent on the ECA server, where corresponding animation is generated, and then streamed in the

form of the audio/video stream to the user's STB. As seen in Figure 8, XBMC-based interfaces with integrated ECA result in much better personification of the system than e.g. just playing out generated speech to the user.

Several members in the lab evaluated generated synthetic expressions, and performed by ECA EVA in the UMB-SmartTV system. They evaluated lip-sync, symbolic representation of meaningful words, and alignment of movement phases and synthesized speech. Evaluators agreed that speech and visual pronunciation are in acceptable temporal sync, and 35% of them suggested to further improve the correlation between visual and audio stressing. 55% of the observed sequences adequately represented the verbal content and 30% of the sequences were observed as a meaningful word mismatch. Nevertheless, when the meaningful word was suggested to them, most of them agreed that the representation is adequate, and appears quite natural. 15% of the sequences were evaluated as out-of-sync in either preparation, and/or stroke movement phase. In this case the observed movement either started or stopped slightly premature. The evaluation showed that most of the generated behavior of the proposed system can be assessed as viable, and close to human-like, and very important feature for future smart TV units.

VII. Conclusion And Future Research

This paper presented a TTS-driven non-verbal behavior system for co-verbal gesture synthesis for UMB-SmartTV system. Its architecture and algorithm used to symbolically and temporarily synchronize the non-verbal expressions with verbal information were presented in detail. Further, we have presented how meaningful parts of verbal content are determined and selected based on word-type-order and semiotic patterns, how a visual representation of meaning can be selected, how the structure of its propagation can be generated as a sequence movement-phases (based on lexical affiliation and semiotic rules), and how movement-phases and durations of movements can be aligned with the verbal content. Finally, the procedural script is formed that can be used for driving synthetically generated synchronized verbal and non-verbal behavior. The produced synthetic behavior reflects very high-degree of lip-sync and iconic, symbolic, and indexical expressions, as well as adaptors, and most of the generated behavior appears very 'natural', and may adequately represent the verbal content.

In our future work investigation will be oriented towards expressive TTS-models in order to take advantage of animating affective ECAs. Further, in order to improve the rules stored within semiotic grammar we will further annotate video corpora, fine tune existing rules (especially regarding the movement dynamics), and create additional shapes (representing meaning of words and word phrases). Our goal is also to further enrich gestural dictionary. Namely, by annotating additional segments of video corpora we will be able to create lots of additional gesture-instances. This will most certainly contribute to the diversity (that is typical for naturalness) and expressive capabilities of ECAs.

References

[1] SoonChoul Kim, Bumsuk Choi, Youngho Jeong, Jinwoo Hong, Jinwook Chung. 2012. An architecture of augmented broadcasting service for next generation smart TV. IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), pp.1-4, 27-29, doi: 10.1109/BMSB.2012.6264289.

[2] Sorwar, G., and Hasan, R. 2012. Smart-TV Based Integrated E-health Monitoring System with Agent Technology. In Proc. of 26th International Conference on Advanced Information Networking and Applications Workshops (WAINA), pp.406-411, doi: 10.1109/WAINA.2012.155.

[3] Pyung-Soo Kim, and Soo Ho Ahn. 2011. A home-oriented IPTV service platform on residential gateway. In Proc. of 8th International Conference on Information, Communications and Signal Processing (ICICS), pp.1-5, doi: 10.1109/ICICS.2011.6174245.

[4] Regina Bernhaupt and Katherine Isbister. 2013. A new perspective for the games and entertainment community. In Proc. of Extended Abstracts on Human Factors in Computing Systems (CHI EA '13). ACM, New York, NY, USA, pp. 2489-2492. DOI=10.1145/2468356.2468812. http://doi.acm.org/10.1145/2468356.2468812.

[5] Trisha T. C. Lin. 2013. Convergence and regulation of multi-screen television: The Singapore experience. Telecommunications Policy, vol. 37, no. 8, pp. 673-685.

[6] Furuta, S., Kawashima, K., Otsuka, T., Yamaura, T., Otsuka, R. 2012. The development of the voice read-out system for digital television receiver. IEEE 1st Global Conference on Consumer Electronics (GCCE), pp. 461-463, doi: 10.1109/GCCE.2012.6379658.

[7] Schröder, M. 2011. The SEMAINE API: a component integration framework for a naturally interacting and emotionally competent embodied conversational agent. PhD Thesis. http://scidok.sulb.uni-saarland.de/volltexte/2012/4544.

[8] Bernhaupt R., and Isbister K. 2012. Games and entertainment community SIG: shaping the future. In Proc. of CHI '12 Extended Abstracts on Human Factors in Computing Systems (CHI EA '12). ACM, New York, NY, USA, pp. 1173-1176. DOI=10.1145/2212776.2212416.

[9] De Carolis B., Mazzotta I., Novielli N., and Pizzutilo, S. 2013. User Modeling in Social Interaction with a Caring Agent. In book: User Modeling and Adaptation for Daily Routines. Martín, Estefanía, Haya, Pablo A., and Carro, Rosa M. (eds.). Human-Computer Interaction Series, Springer-Verlag London, pp. 89-116.

[10] Doumanis I., and Serengul S. 2013. An empirical study on the effects of embodied conversational agents on user retention performance and perception in a simulated mobile environment. In Proc. of 9th International Conference on Intelligent Environments (IE'13), Athens, Greece, pp. 431-442.

[11] Kunc L., Míkovec Z., and Slavík P. 2013. Avatar and Dialog Turn-Yielding Phenomena. International journal of Technology and Human Interaction, vol. 9, no. 2, pp. 66-88, doi:10.4018/jthi.2013040105.

[12] Thiebaux M., Marsella S., Marshall A.N., Kallmann M. 2008. SmartBody: behavior realization for embodied conversational agents. In Proc. of the 7th international joint conference on Autonomous agents and multiagent systems (AAMAS '08), pp. 151-158.

[13] Kopp S., Wachsmuth I. 2002. Model-based animation of co-verbal gesture. In Proc. of Computer Animation, 2002, pp. 252-257.

[14] Hadar U., Krauss R.K. 1999. Iconic gestures: the grammatical categories of lexical affiliates. Journal of Neurolinguistics, vol. 12, no. 1, pp. 1 - 12.

[15] Straube B., Green A., Bromberger B., Kircher T. 2011. The differentiation of iconic and metaphoric gestures: common and unique integration processes. Human Brain Mapping, vol. 32, no. 4, pp. 520-533.

[16] Kopp S., and Wachsmuth I. 2004. Synthesizing multimodal utterances for conversational agents. Journal of Computer Animation and Virtual Worlds, vol. 15, no. 1, pp. 39-52.

[17] Holler J., Wilkin K. 2011. Co-speech gesture mimicry in the process of collaborative referring during face-to-face dialogue. Journal of Nonverbal Behavior, vol. 35, no. 2, pp. 133-153.

[18] Allwood J. 2010. Dialog Coding - Function and Grammar. Gothenburg Papers in Theoretical Linguistics, 85, 2010.
[19] Grenfell M. 2011. Bourdieu, Language and Linguistics. Continuum International Publishing Group, 2011.
[20] Kamailio. 2013. http://www.kamailio.org/w/, WWW.
[21] Bazot P., Huber R., Kappel J., Subramanian B.S., Oguejiofor E., Georges B., Jackson C., Martin C., Sur A. 2007. Developing SIP and IP Multimedia Subsystem (IMS) Applications, Redbooks, 2007.
[22] Rojc M., Mlakar I. 2009. Finite-state machine based distributed framework DATA for intelligent ambience systems. V: BULUCEA, Cornelia A. (ed.). Recent advances in computational intelligence, man-machine systems and cybernetics: proceedings of the 8th WSEAS International Conference on computational intelligence, man-machine systems and cybernetics (CIMMACS '09), Puerto de la Cruz, Tenerife, Canary Islands, Spain, pp. 80-85.
[23] XBMC. 2013. http://xbmc.org/, WWW.
[24] McNeill D. 1992. Hand and Mind - What gestures reveal about thought. The University of Chicago Press, Chicago.
[25] Vilhjalmsson H., Cantelmo N., Cassell J., Chafai N.E., Kipp M., Kopp S., Mancini M., Marsella S., Marshall A. N., Pelachaud C., Ruttkay Z., Thórisson K.R., van Welbergen H., van der Werf R.J. 2007. The behavior markup language: recent developments and challenges. In Proc. of IVA'07, vol. 4722, pp. 99-111.
[26] Mlakar I., Rojc M. 2011. Towards ECA's Animation of Expressive Complex Behaviour. In book: Analysis of Verbal and Nonverbal Communication and Enactment. A. Esposito, A. Vinciarelli, K. Vicsi, C. Pelachaud, A. Nijholt (eds.). LNCS 6800, pp. 185-198.
[27] Mlakar I., Rojc M. 2012. Capturing form of non-verbal conversational behavior for recreation on synthetic conversational agent EVA. WSEAS Trans. Comput. [Print ed.], vol. 11, no. 7, pp. 218-226.
[28] Rojc M., Kačič Z. 2007. Time and Space-Efficient Architecture for a Corpus-based Text-to-Speech Synthesis System. Speech Communication, vol. 49, no. 3, pp. 230-249.
[29] Mlakar I., Rojc M. 2010. Personalized expressive embodied conversational agent EVA. V: MASTORAKIS, Nikos E., MLADENOV, Valeri (eds.). Advances in visualization, imaging and simulation: proceedings of the 3rd WSEAS International Conference on visualisation, imaging and simulation (VIS '10), University of Algarve, Faro, Portugal, [S. l.]: WSEAS Press, pp. 123-128.
[30] Pine K., Bird H., Kirk E. 2007. The effects of prohibiting gestures on children's lexical retrieval ability. Developmental Science 10, pp. 747-754.
[31] Mlakar I., Rojc M. 2012. Recreation of spontaneous non-verbal behavior on a synthetic agent EVA. V: RUDAS, Imre J. (ed.). Recent researches in artificial intelligence and database management: proceedings of the 11th WSEAS International conference on Artificial intelligence, knowledge engineering and data bases (AIKED '12), Cambridge, UK, [S. l.]: World Scientific and Engineering Academy and Society and Society Press, cop. 2012, pp. 225-230.
[32] Kita S., van Gijn I., van der Hulst H. 1998. Movement phases in signs and co-speech gestures, and their transcription by human coders. In: Gesture and sign language in human-computer interaction. I. Wachsmuth M. Frohlich (eds.), pp. 23-35.
[33] Loehr D. 2004. Gesture and intonation, Doctoral Dissertation, Georgetown University.
[34] XBMC, Open Source Home Theatre Software. 2013. http://xbmc.org/, last visited in July, 2013.
[35] Arnaud J., Négru D., Sidibé M., Pauty J., and Koumaras H. Adaptive IPTV services based on a novel IP Multimedia Subsystem. Multimedia Tools and Applications, vol. 55, no. 2, 2011, pp. 333-352.
[36] Mlakar I., Rojc M. 2011. EVA: expressive multipart virtual agent performing gestures and emotions. International journal of mathematics and computers in simulation, vol. 5., no. 1, pp. 36-44.
[37] Poggi I., Pelachaud C., de Rosis F., Carofiglio V., de Carolis B. 2005. Greta. A believable embodied conversational agent. In book: Multimodal Intelligent Information Presentation. Oliviero Stock and Massimo Zancanaro (eds.), vol. 27, pp. 3-25.
[38] Poggi I., Pelachaud C. 2000. Performative facial expressions in Animated faces. In book: Embodied conversational agents. Justine Cassell, Joseph Sullivan, Scott Prevost, Elizabeth Churchill (eds.), MA, USA, pp. 155-188.
[39] Ball, G., Breese, J. 2000. Emotion and personality in a conversational agent. In book: Embodied conversational agents. Justine Cassell, Joseph Sullivan, Scott Prevost, Elizabeth Churchill (eds.), MA, USA, pp. 189-219.
[40] Heloir A., Kipp M. 2009. EMBR—a realtime animation engine for interactive embodied agents. In Proc. of the 9th International Conference on Intelligent Virtual Agents (IVA '09), Springer, Amsterdam, The Netherlands, pp. 393–404.
[41] DeCarolis B., Pelachaud C., Poggi I., Steedman M. 2004. APML, a mark-up language for believable behavior generation. In H. Prendinger and M. Ishizuka (eds.), Life-like Characters. Tools, Affective Functions and Applications, Springer-Verlag Berlin Heidelberg, pp. 65-85.
[42] Vilhjalmsson, H., Cantelmo N., Cassell J., Chafai N., Kipp M., Kopp S., Mancini M., Marsella S., Marshall A., Pelachaud C., Ruttkay Z., Thorisson K., van Welbergen H., van der Werf R. 2007. The Behavior Markup Language: Recent Developments and Challenges. IVA 2007, LNAI 4722, Springer-Verlag Berlin Heidelberg, pp. 99-111.
[43] Goslin, M., Mine, M. R. 2004. The Panda3D Graphics Engine, Computer, vol. 37, no.10, pp.112-114.
[44] Ekman, P. Darwin. 2003. Deception, and Facial Expression. Annals of the New York Academy of Sciences, vol. 1000, no. 1, pp. 205-221.