INTERNATIONAL JOURNAL OF COMPUTERS AND COMMUNICATIONS
DOI: 10.46300/91013.2021.15.1 Volume 15, 2021

TTS-driven Embodied Conversation Avatar for UMB-SmartTV

Matej Rojc
Faculty of Electrical Engineering and Computer Science,
University of Maribor,
Smetanova ulica 17
Maribor, Slovenia
matej.rojc@uni-mb.si

Marko Presker
Faculty of Electrical Engineering and Computer Science,
University of Maribor
Smetanova ulica 17
Maribor, Slovenia
marko.presker@uni-mb.si

Zdravko Kačič
Faculty of Electrical Engineering and Computer Science,
University of Maribor
Smetanova ulica 17
Maribor, Slovenia
kacic@uni-mb.si

Izidor Mlakar
Panevropa d.o.o,
Ob Ribniku 24
Maribor, Slovenia
izidor@panevropa.com

Received: January 20, 2021. Revised: April 2, 2021. Accepted: April 8, 2021. Published: April 14, 2021.

Abstract— **When human-TV interaction is performed by remote controller and mobile devices only, the interactions tend to be mechanical, dreary and uninformative. To achieve more advanced interaction, and more human-human like, we introduce the virtual agent technology as a feedback interface. Verbal and co-verbal gestures are linked through complex mental processes, and although they represent different sides of the same mental process, the formulations of both are quite different. Namely, verbal information is bound by rules and grammar, whereas gestures are influenced by emotions, personality etc. In this paper a TTS-driven behavior generation system is proposed for more advanced interface used for smart IPTV platforms. The system is implemented as a distributive non-IPTV service and integrated into UMB-SmartTV in a service-oriented fashion. The behavior generation system fuses speech and gesture production models by using FSMs and HRG structures. Features for selecting the shape and alignment of co-verbal movement are based on linguistic features (that can be extracted from arbitrary input text), and prosodic features (as predicted within several processing steps in the TTS engine). At the end, the generated speech and co-verbal behavior are animated by an embodied conversational agent (ECA) engine and represented to the user within the UMB-SmarTV user interface.**

Keywords— embodied conversational agents, Smart TV, gestures, HCI interface, 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 generating speech from 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]. ECAs may further personificate smart TV platforms [9, 10] and may have a huge impact on interactivity and personalization of IPTV and other services provided by smart TV platforms.

The main reason for integration of non-verbal modalities together with text-to-speech systems is to better emulate the natural course of the dialogue. Such integration enables users to interact with a virtual person, and makes them feel more comfortable when "talking" to a smart TV UI. The second reason is hidden in issues that generally occur whilst using human-machine interaction systems: repentance 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 repentance 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.

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 voice. 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 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 rasperry pi platform and Z-wave mesh networking technology. In the next section, multimodal platform integrating virtual agent technology is presented into detail.

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: automatic text-to-speech engine (TTS), automatic 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.

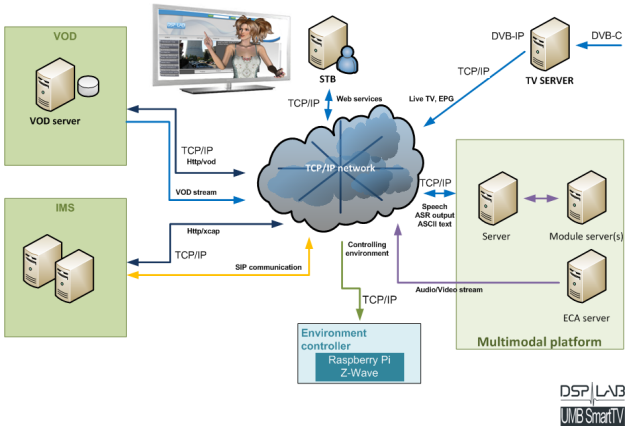

Fig. 1. Architecture of the UMB-SmartTV.

The spoken dialog manager engine drives the interaction between users and UMB-SmartTV. 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. 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 directly 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. How this fusion of core TTS system and virtual agent technology is performed, will be presented in the next section into 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 of non-verbal behavior involve semiotic grammar, gesture dictionary, and lexical affiliation that are included into the behavior generation system as external resources. 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]).

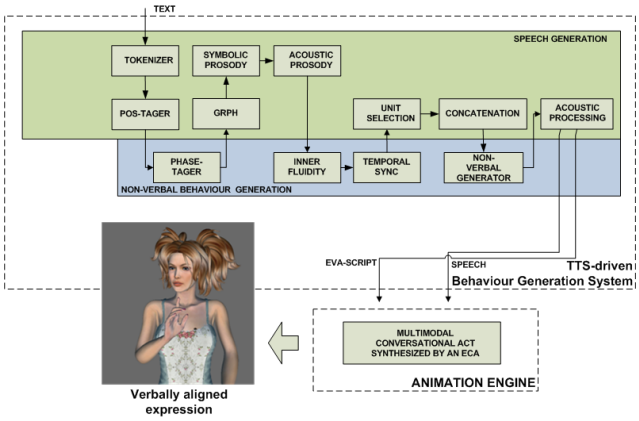

Fig. 2. The TTS-driven behavior generation system.

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 an embodied conversational agent in 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 is 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's output generation is very 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 will be presented into more 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 prepared gesture lexicon (gesture affiliates). Previous modules (POS-tagger) provide the needed input information. All data extracted from input text are stored in the heterogeneous relation graphs (HRG) (Figure 3), where they can be easily find and accessed, and transferred between several modules in the system.

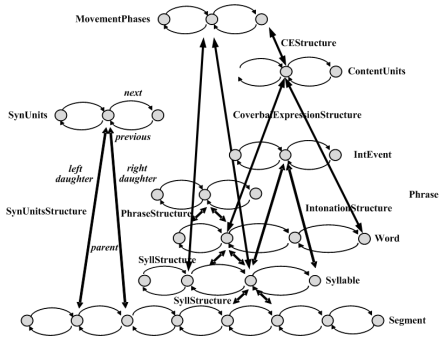

Fig. 3. The HRG structure for storing verbal and non-verbal information within behavior generation system.

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 matched 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. Namely, 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]).

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 common 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 phase tagging deque) are compared against prosodic word phrases (logical content segments), as predicted in the symbolic prosody deque (module generally used for text-to-speech synthesis). Namely, these prosodic phrases indicate 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 is "searching of emphasised words" and 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 core TTS system (by using CART models) and assigned to the 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 behavior generation system, by using temporal information, the duration of each movement phase can also be determined. The deque draw 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. Additionally, it 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 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 within 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 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 at the end 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).

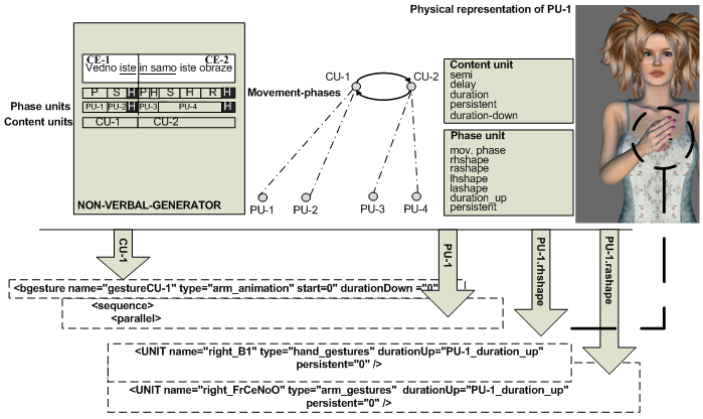

Fig. 4. Transformation of data in the HRG structure into procedural animation specification.

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].

V. 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). In the on-line system they can be automatically selected and also temporally adjusted during the presented 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 5, 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.

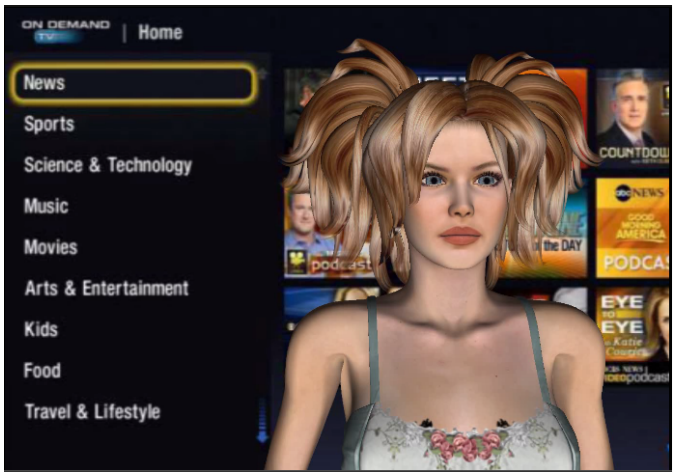

Fig. 5. ECA integrated in the XBMC interface of the UMB-SmartTV system.

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.

VI. 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. Broadband Multimedia Systems and Broadcasting (BMSB), 2012 IEEE International Symposium, pp.1-4, 27-29 June 2012, doi: 10.1109/BMSB.2012.6264289.

[2] Sorwar, G., Hasan, R. Smart-TV Based Integrated E-health Monitoring System with Agent Technology. Advanced Information Networking and Applications Workshops (WAINA), 2012 26th International Conference, pp.406-411, 26-29 March 2012, doi: 10.1109/WAINA.2012.155.

[3] Pyung-Soo Kim, Soo Ho Ahn. A home-oriented IPTV service platform on residential gateway. Information, Communications and Signal Processing (ICICS) 2011 8th International Conference, pp.1-5, 13-16 Dec. 2011, doi: 10.1109/ICICS.2011.6174245.

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

[5] Trisha T. C. Lin. Convergence and regulation of multi-screen television: The Singapore experience. Telecommunications Policy, 37(8):673–685, September, 2013.

[6] Furuta, S., Kawashima, K., Otsuka, T., Yamaura, T., Otsuka, R. The development of the voice read-out system for digital television receiver. Consumer Electronics (GCCE), 2012 IEEE 1st Global Conference, pp. 461-463, 2-5 Oct. 2012, doi: 10.1109/GCCE.2012.6379658.

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

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

[9] Carolis, Berardina and Mazzotta, Irene and Novielli, Nicole and Pizzutilo, Sebastiano . 2013. User Modeling in Social Interaction with a Caring Agent. In book User Modeling and Adaptation for Daily Routines, Ed. Martín, Estefanía and Haya, Pablo A. and Carro, Rosa M., Human–Computer Interaction Series, Springer, London, pp. 89-116.

[10] Doumanis, Ioannis and Smith, Serengul. 2013. An empirical study on the effects of embodied conversational agents on user retention performance and perception in a simulated mobile environment. In: The 9th International Conference on Intelligent Environments (IE'13), 16-17July 2013, Athens, Greece.

[11] Kunc, Ladislav, Zdenek Míkovec and Pavel Slavík. 2013. Avatar and Dialog Turn-Yielding Phenomena. IJTHI 9.2 (2013): pp. 66-88. Web. 31 Jul. 2013. doi:10.4018/jthi.2013040105.

[12] M. Thiebaux, S. Marsella, A.N. Marshall, M. Kallmann, SmartBody: behavior realization for embodied conversational agents. In proc. of AAMAS '08 (2008), pp. 151-158.

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

[14] U. Hadar, R.K. Krauss, Iconic gestures: the grammatical categories of lexical affiliates, J. of Neurolinguistics 12(1), 1999, pp. 1 - 12.

[15] B. Straube, A. Green, B. Bromberger, T. Kircher, The differentiation of iconic and metaphoric gestures: common and unique integration processes, Hum Brain Mapp. 32(4), 2011, pp. 520-33.

[16] S. Kopp, I. Wachsmuth. Synthesizing multimodal utterances for conversational agents, J. of Comput Animat. Virtual Worlds 15(1), 2004, pp. 39–52.

[17] J. Holler, K. Wilkin, Co-speech gesture mimicry in the process of collaborative referring during face-to-face dialogue, J. of Nonverbal Behavior 35, 2011, pp. 133-153.

[18] J. Allwood, Dialog Coding – Function and Grammar. Gothenburg Papers in Theoretical Linguistics, 85, 2010.

[19] G. Michael, Bourdieu, Language and Linguistics, Continuum International Publishing Group, 2011.

[20] Kamailio, http://www.kamailio.org/w/, WWW.

[21] P. Bazot, R. Huber, J Kappel, B.S: Subramanian, E. Oguejiofor, B. Georges, C. Jackson, C. Martin, A. Sur, Developing Sip and IP Multimedia Subsystem (IMS) Applications, 2007.

[22] M. Rojc, I. Mlakar. Finite-state machine based distributed framework DATA for intelligent ambience systems. V: BULUCEA, Cornelia A. (ur.). 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, December 14-16, 2009, (Electrical and computer engineering series). [S. l.]: WSEAS Press, cop. 2009, pp. 80-85.

[23] XBMC, http://xbmc.org/, WWW.

[24] D. McNeill, Hand and Mind - What gestures reveal about thought, The University of Chicago Press, Chicago, 1992.

[25] H. Vilhjalmsson, N. Cantelmo, J. Cassell, N.E. Chafai, et al., The behavior markup language: Recent developments and challenges, In proc. of IVA'07 (2007).

[26] I. Mlakar, M. Rojc. Towards ECA's Animation of Expressive Complex Behaviour, In: A. Esposito, A. Vinciarelli, K. Vicsi, C. Pelachaud, A. Nijholt (Eds.), Analysis of Verbal and Nonverbal Communication and Enactment. The Processing Issues, LNCS 6800, 2011, pp. 185-198.

[27] I. Mlakar, M. Rojc. Capturing form of non-verbal conversational behavior for recreation on synthetic conversational agent EVA. WSEAS Trans. Comput.. [Print ed.], 2012, vol. 11, iss. 7, pp. 218-226.

[28] M. Rojc, Z. Kačič, Time and Space-Efficient Architecture for a Corpus-based Text-to-Speech Synthesis System, Speech Communication 49(3), 2007, pp. 230-249.

[29] I. Mlakar, M. Rojc. Personalized expressive embodied conversational agent EVA. V: MASTORAKIS, Nikos E. (ur.), MLADENOV, Valeri (ur.). 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, November 3 - 5, 2010. [S. l.]: WSEAS Press, 2010, pp. 123-128.

[30] K. Pine, H. Bird, E. Kirk, The effects of prohibiting gestures on children's lexical retrieval ability, Developmental Science 10, 2007, pp. 747–754.

[31] I. Mlakar, M. Rojc. Recreation of spontaneous non-verbal behavior on a synthetic agent EVA. V: RUDAS, Imre J. (ur.). 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, February 22-24, 2012. [S. l.]: World Scientific and Engineering Academy and Society and Society Press, cop. 2012, pp. 225-230.

[32] S. Kita, I. van Gijn, H. van der Hulst, Movement phases in signs and co-speech gestures, and their transcription by human coders, In: I. Wachsmuth M. Frohlich (Eds.), Gesture and sign language in human-computer interaction (1998), pp. 23–35.

[33] D. Loehr, Gesture and intonation, Doctoral Dissertation, Georgetown University, 2004.

[34] XBMC, Open Source Home Theatre Software, http://xbmc.org/, last visited in July, 2013.

[35] J. Arnaud, D. Négru, M. Sidibé, J. Pauty, and H. Koumaras, "Adaptive IPTV services based on a novel IP Multimedia Subsystem", Multimedia Tools and Applications, vol. 55, no. 2, 2011, pp. 333-352.

[36] I. Mlakar, M. Rojc. (2011). EVA: expressive multipart virtual agent performing gestures and emotions. International journal of mathematics and computers in simulation, vol. 5., iss. 1, pp. 36-44.