Jurnal Teknologi

Full paper

A Natural Conversational Virtual Human with Multimodal Dialog System

Itimad Raheem Ali,[a,*] Ghazali Sulong,[a] Ahmad Hoirul Basori,[b]

[a]*Faculty of Computing, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor Malaysia*
[b]*Interactive Media and Human Interface Lab, Department of Informatics, Faculty of Information Technology, Institut Teknologi Sepuluh Nopember Surabaya, Indonesia*

*Corresponding author: raitimad2@utm.my

Article history

Received : 31 July 2014
Received in revised form :
23 November 2014
Accepted : 1 December 2014

Abstract

The making of virtual human character to be realistic and credible in real time automated dialog animation system is necessary. This kind of animation carries importance elements for many applications such as games, virtual agents and movie animations. It is also considered important for applications which require interaction between human and computer. However, for this purpose, it is compulsory that the machine should have sufficient intelligence for recognizing and synthesizing human voices. As one of the most vital interaction method between human and machine, speech has recently received significant attention, especially in avatar research innovation. One of the challenges is to create precise lip movements of the avatar and synchronize it with a recorded audio. This paper specifically introduces the innovative concept of multimodal dialog systems of the virtual character and focuses the output part of such systems. More specifically, its focus is on behavior planning and developing the data control languages (DCL).

Keywords: Speech Synchronization; Dialog Behavior Systems.

1.0 INTRODUCTION

In the last decade, there has been an increasing interest in real time virtual character designing. This interest has not only been limited to games or online communication, but also for dialog based systems such as infotainment applications and tutoring systems. This is considered as the major challenge to human computer interface technologies in general and virtual reality concepts in particular. The main aim was to develop an intuitive interface instead of the GUI which did not improve for more than four decades. However, virtual characters are considered better as an interactive interface to stimulate facial expressions and communicative behaviors.

Here, the main challenge is to create a connection between low level graphics and high level behavior control. Hence, a gap still exists between realization and the expression behavior planning. Therefore, it is essential that computer scientists and psychologists work in collaboration for synthesizing multilayer personality models for animation using conversational virtual characters.

2.0 SPEECH SYNCHRONIZATION

In an animation, the synthesized virtual human character shall become a natural actor if engaged in a conversation where a correct synchronization is created between speech and lip movement. A human observer is always sensitive to facial expressions. It can detect the smallest inaccuracies concerning lip movements and the face in relation to the uttered speech.

Speech animation is necessary for automation as its implementation is a laborious one. Its main purpose is for obtaining optimal series of animation parameters associated with certain speech. So, the relationship between a set of phonemes and the necessary animation parameters does not follow a linear function. The reason is that identical speech can be produced in different ways. It depends on the emotion involved, although, a high correlation exists between lip motion and phonemes.

The animation imposed on a virtual character is actually dependent on the input to the process of synthesis. However, in many cases, the input contains a split, concatenated or an incomplete text. So, the dominance functions are used and trajectories of the input are generated to produce the desired animation. But, in some cases, parameterized speech is directly used as an input. This will require the format analysis, linear regression or probabilistic modeling for generating an appropriate motion. The animation creator will possibly use a text based approach when a spoken text is available. In digital games or movies applications, where the animation moves faster and in real time, the speech based input is considered more suitable for rapid prototype work.

3.0 MULTIMODAL DIALOG SYSTEMS

The goal of Multimodal dialog systems (MDS) is to enhance the interaction between human and compute with respect to the realism, robustness, understandable, efficiency and enjoyment. Therefore, Multimodal dialog systems was extend traditional speech based dialog systems with added modalities for input(pen, camera, multi-touch, sensors) and output (3D graphics, robots, animations, physical artifacts,)[1] .Actually, virtual character was considered as an output device for an MDS cause allowing to symmetric interaction[2].

Virtual characters can utilize verbal and nonverbal communication like facial expressions, eye behavior, gestures, emotions, and postures. This is possible and a challenge at the same time cause unnatural behavior has the reverse effect of the users or may be reject it. This is referred as uncanny valley's effect, was made more human-like in its behaviors, motions and emotion response.

The shift from language generation in a natural way to generating a multi-modal behavior is the control components of Multimodal dialog systems (MDS). The role of dialog system is to generate voice and to synchronize the facial expressions. Generally, multimodal signals have meaning and communicative function on one hand and their visible behavior on the other. For instance, a deictic meaning like "here" and "there" shows a deictic pointing gesture. Modalities are interconnected not only meaning wise but also in respect to their temporal organization. For example, we hypothesize gesture to co-occur with the corresponding word or phrase in speech (known as the lexical affiliate), even though, the question still debatable[3]. To observe relevant terms that is necessary to understand the multimodal dialog system, we will define one example on it, an embodied conversational agent (ECA), it is a virtual agent that has artificial intelligent, cognitive and expressive capabilities to simulate human capabilities (social interactive and independence) communicating with the users through verbal and nonverbal expressions[4].

Dealing with the embodied conversational agents (ECA) requires multidisciplinary participation between different fields such as AI, computer graphics, psychology, social science, and linguistics, cognitive, and so on so forth. The main issues of a computer graphics are characters modeling, automatic rendering, dynamic simulations, and realistic animations. One focus of this article is to clarify the connecting steps between low-level graphics on the one hand and high level behavior control on the other.

4.0 DIALOG BEHAVIORS CONTROL

The behavior of the virtual character by default contains information about the expressivity and the contents of the communicative act. It is determined by the communicative intention and by the character's fundamental general behavior tendency. The behavior is an essential aspect of communicative acts; therefore modeling of communicative behavior must be handled earlier at a higher level must be able to display a flexible behavior.

The need in higher level interfaces which allows abstract definition and control of the object behavior needs better interactive control and scheduling mechanisms. Hence, in SAIBA framework[5] for interactive agents, three stages of the behavior generation were discovered which are usually independent of the concrete realization of a character known as intent planning, behavior planning and behavior realization. When the intent planning is focused, many approaches can be identified which are reviewed in this section. For defining the problem, based on the human perceived and the virtual world on one hand and the goals of the virtual characters on the other hand, the intent planning module takes the decision about what actions is needed to be taken next. Such type of action includes high level behaviors such as talking a simple sentence or walk to a target place. On the other hand, lower level behaviors such as producing a gesture or to change posture fall under the responsibility of the behavior planner. Actually, the realization of the actions is not performed by the intent planner. Instead of it, it can be assumed that the output actions can be formulated on abstract level by using BML language[6]. It can also be executed through dedicated realization engines such as speech synthesizer and character animation engine. There are many famous approaches concerned with how to control the behavior of a virtual character, but we will focus on the following:

4.1 State-Based System Approach

This approach involves representing and visualizing the character's mind by states and transitions. As the graph is traversed, actions attached to state or transition is executed. This approach has been previously used successfully in CSLU Toolkit for speech-based interaction[7]. In Crosstalk and COHIBIT systems, the so called sceneflow control the interactive embodied agents. It is an extended hierarchical finite state machine (FSM) where an atomic state is represented by a rod and super node that contains another FSM[8]. The transitions may be conditional, probabilistic, or interruptive (for exciting supernode). These nodes and the edges might have pre-scripted scenes attached which may specify the dialogue and the non-verbal actions. Parallel Transition Networks (PaT-Nets) is also a concept of similar type which has facilities for parallelism. These were used in character animation for the combined control of high level behavior and the low level animation[9].

4.2 Connectionist Approaches

Several scientists experimented with biologically motivated recognition methods that use an interconnected network of simple units. Percepts from the environment are fed into the input units and propagated through the network to the output layer in which the most active unit represents a decision, such as the Norms in the Creatures computer game are controlled by two neural networks, one for decision making and attention, and one for selecting sensory motor commands. In another system an autonomous virtual human is controlled by Tyrell's free-flow hierarchies, an ethologically motivated network architecture[10].

4.3 Multi-Agent Systems Architectures

Multi-agent systems research suggests several concrete architectures for controlling intelligent agents. BDI (Belief-Desire-Intention) is a cognitively motivated architecture where beliefs represent information about the world and desires are options available to the character and intentions denote goals that an agent is committed to. A planner is usually used to generate a sequence of actions based on the current set of intentions and beliefs. BDI was employed in agent scenarios modeling autonomous life-like behaviors and social interaction behaviors[11]. Complex behavior is decomposed into simple condition-action behaviors and organized into layers. If percepts from the environment satisfy conditions of several behaviors, the lowest one is selected. Stefan Scherer at (2013) presented a novel multimodal corpus recorded with a virtual audience public speaking training platform, to utilize the description of behavior to automatically approximate the overall assessment of the performance using support vector regression in a speaker-independent experiment and approaching human performance[12].

4.4 Hybrid Approaches

For balancing the needs of reactive and deliberative behaviour, the hybrid architecture combines many control methods. For example, the REA agent processes inputs either by a set of hardwired reactions which leads to an immediate input (i.e. the gaze of agents tracks the user's movement) or through a deliberative planning based module (to select utterances as per a communicative goal[13].

The Cross Talk System compiles a hierarchical FSM into plan operators that can be utilized in a classical plan based approach at runtime and the ITeach system runs an FSMand a RBS in parallel. It synchronizes them by using shared variables[10]. Li et al (2013) presented a hybrid method for synthesizing natural animation of facial expression with data from motion capture. The captured expression was transferred from the space of source performance to that of a 3D target face using an accurate mapping process in order to realize the reuse of motion data. The transferred animation was then applied to synthesize the expression of the target model through a framework of two-stage deformation[14]. When scenarios become complex, hybrid approaches are often necessary. By using appropriate technology, different aspects can be handled. For example, through RBS, knowledge processing aspects can be effectively handled. Procedural aspects can be best modeled through FSM or state chart (ideally with a graphical interface).

■5.0 CONTROL LANGUAGES

Various kinds of behavioral models are dependent on the level of autonomy of the character and on whether body and mind are considered independent or not. Control languages serve as reusable representation of agent behavior and separation between modules that implement different functions, for example, behavior planning and realization[10]. The BEAT toolkit for the automated generation of nonverbal behavior used multiple languages to pass information from module to module[15]. The SAIBA model emerged out of this framework where the functional markup language (FML) is used for encoding the communicative intent without referring to physical realization and the behavior markup language (BML) specifies the verbal utterance and nonverbal behaviors like gesture, posture and facial expression[16].

Defining an additional dictionary of behavior descriptions, the "Gesticon"[10], the language distinguishes between abstract behavior definitions and concrete realizations. MURML[17] and APML[18] are, like BML, specification languages for physical realization. MURML allows describing gestures by defining spatiotemporal constraints and sub movements of a gesture stroke. An application example is demonstrated with the anthropomorphic agent Max. MPML/MPML3D was designed for web-based scenarios and codes verbal and nonverbal behavior, presentation flow and integration with external objects. VHML is an XML-based language which consists of several sub-languages for describing the character, like GML for its gestures, FAML for facial animation, BAML for body animation, EML for emotions, etc[10].

Emotion Markup Language (EML) was designed for representing emotional states for stimulating a user interface or of a human user in standardized way. To describe an emotion-related behaviour with emotion ML are available http://www.w3.org/TR/emotionml/\#s5.1.3. Languages like BML use concepts like relative timing and lexicalized behaviors;[5] outlined the necessity for an additional declarative animation layer, a thin wrapper around the animation engine and are situated below higher-level behavior control layers for abstracting away from implementation details while giving access to the functionality of the engine.

To develop interactive virtual humans on the graphic side, it is necessary to take into account not only the geographic model and some basic ways of animating but also the aspects related to various level of abstraction. Hence, in[1], the authors propose a generic and layered software architecture which focuses the behavioral aspects, while providing animation models which include collision detection and path planning. The digital production of realistic eye movement will require an emotional eye movement animation scripting tool such as markup language (EEMML) which was developed by Zheng Li et al (2011). Salvati et al (2011) improved the animation scripting tool of Zheng Li et al when he created (FSM) Face image Synthesis Module which is a general toolkit for building an easily customize embodied agent based on multimodal integrated dialog, speech synthesis, speech recognition and face image synthesis[19].

Human Markup Language (HML) supports the modeling of human behavior at the very high level. HML is backed-up by the Internet repository system and a set of tools, that tags various verbal and non-verbal communication cues used in human-to-human interactions. Examples of languages supporting the channel mixing concept include VHML, SMILE-AGENT, and RRL[20].

The Web application domain has brought its own set of XML-based languages that help Web page designers enhance human-machine interaction experience. Multimodal Presentation Mark-up Language (MPML) builds on the body of the Microsoft Agent to create predefined animation sequences. Behavior Expression Animation Toolkit (BEAT)[15] processes the XML input description in the form of a tree containing both verbal and non-verbal signals, to produce the synchronized animated sequence on the output. Kunc and Kleindienst were presented an architecture and interaction language ECAF, which we used for authoring several ECA-based applications[20].

■6.0 CONCLUSION

The process of synchronizing input speech with synthesized facial motion is known as lip-synching or speech motion synthesis. In speech synthesis, input speech is normally represented by a standard speech unit called phoneme. The

conversion from speech to phoneme can be done manually (Parke, 1975) or automatically (Lewis, 1991; Bregler et al. 1997). However, it is acceptable if the animated dialogue can interconnect every important expressions of emotions such as the correct movements of the eyes (especially the timing of a blink). The synchronized appearance of a smile (especially the hint and the duration of the smile) and the correct facial expressions (such as the radiant look of happiness, the frown from sadness, the fierce look of anger and the fearful look of humiliation).

A conversation or dialog can be simulated in many different ways by different emotional states are identified and encoded. The lexicon obtained should contain enough emotion variables, so that, different personalities required in a dialogue can be displayed. Voice intonation and talking speed are normally controlled, so that the simulated talking head can express sufficient emotional expressiveness relevant to its character in the animation video. The development of modules of dialog systems containing integration of voice intonation and correct talking speed is an important future research topic towards the synthesis of human-like talking head. In the synthesis of virtual human character, it is necessary to observe and listen beyond the words, so that the correct sense of emotions can be captured in the near future works.

Acknowledgement

Authors are grateful to Universiti Teknologi Malaysia, Research Management Centre (RMC) of UTM and Department of Informatics, Faculty of Information Technology, Institut Teknologi Sepuluh Nopember Surabaya, Indonesia for financial and technical supports.

References

[1] G. Zoric, R. Forchheimer, and I.S. Pandzic. 2010. *On creating multimodal virtual humans-real time speech driven facial gesturing, Multimedia Tools and Applications*. 54(1): 165–179.

[2] W. Wahlster. 2006. *Dialogue Systems Go Multimodal The SmartKom Experience*. Springer Berlin Heidelberg. 3–27.

[3] G. Ferré. 2010. *Timing Relationships between Speech and Co-Verbal Gestures in Spontaneous French*.

[4] A. Cerekovic and I. S. Pandžic. 2011. Multimodal behavior realization for embodied conversational agent, *Multimedia Tools and Applications*. 54(1): 143–164.

[5] S. Kopp, B. Krenn, S. Marsella, A. Marshall, C. Pelachaud, H. Pirker, K. Thórisson, and H. Vilhjálmsson. 2006. Towards a common framework for multimodal generation: The behavior markup language. in *Intelligent Virtual Agents*. 4133: 205–217.

[6] P. Aggarwal and D. Traum. 2011. *The BML Sequencer: A Tool for Authoring Multi-character Animations*. 428–430.

[7] S. Sutton, R. Cole, J. De Villiers, J. Schalkwyk, P. Vermeulen, M. Macon, Y. Yan, E. Kaiser, B. Rundle, K. Shobaki, P. Hosom, A. Kain, J. Wouters, D. Massaro, and M. Cohen. 1998. *Universal Speech Tools: The Cslu Toolkit*.

[8] G. Skantze and S. Al Moubayed. 2012. *IrisTK: a Statechart-based Toolkit for Multi-party Face-to-face Interaction*. 69–76.

[9] F. López-Colino and J. Colás. 2012. *Spanish Sign Language synthesis system*. Journal of Visual Languages & Computing. 23(3). 121–136.

[10] Y. Jung, A. Kuijper, D. Fellner, M. Kipp, J. Miksatko, J. Gratch, and D. Thalmann. 2011. *Believable Virtual Characters in Human-Computer Dialogs*.

[11] B. Krenn, C. Pelachaud, H. Pirker, and C. Peters. 2011. *Emotion-Oriented Systems*. 389–415.

[12] S. Scherer. 2013. *Towards a Multimodal Virtual Audience Platform for Public Speaking Training*. International Conference on Intelligent Virtual Agents.

[13] J. Cassell and E. Mbodied. 2000. *Human Conversation as a System Framework: Designing Embodied Conversational Agents*.

[14] B. Li, Q. Zhang, D. Zhou, and X. Wei. 2013. *Facial Animation Based on Feature Points*. 11(3).

[15] J. Cassell, H.H. Vilhjálmsson, and T. Bickmore. 2001. *BEAT: the Behavior Expression Animation Toolkit. In Proceedings Of The 28th Annual Conference On Computer Graphics And Interactive Techniques*. 137: 477–486.

[16] S. Kopp, B. Krenn, S. Marsella, and A. N. Marshall. 2011. *Towards a Common Framework for Multimodal Generation: The Behavior Markup Language*.

[17] L.Q. Anh and C. Pelachaud. 2011. *Expressive Gesture Model for Humanoid Robot*. 224–231.

[18] E. Bevacqua, T. Paristech, C. T. Paristech, J. Looser, and C. Pelachaud. 2011. *Cross-Media Agent Platform*. 1(212): 11–20.

[19] M. Salvati and K. Anjyo. 2011. *Developing Tools for 2D / 3D Conversion of Japanese Animations*. 4503.

[20] L. Kunc and J. Kleindienst. 2007. *ECAF: Authoring Language for Embodied Conversational Agents*. Springer-Verlag Berlin Heidelberg. 4629: 206–213

.