Evaluating the spoken dialogue system of a conversational character: A Simulation Study

Yoselie Alvarado[1], Claudia Gatica[2], Veronica Gil Costa[2], and Roberto Guerrero[1]

[1]Laboratorio de Computación Gráfica (LCG)
[2]Laboratorio de Investigación y Desarrollo
en Inteligencia Computacional (LIDIC)
Universidad Nacional de San Luis,
Ejército de los Andes 950
Tel: 02664 420823, San Luis, Argentina
{ymalvarado,crgatica,vgcosta,rag}@unsl.edu.ar

Abstract. Currently, there are many applications like human-computer interactions in which speech technology plays an important role. In particular, embodied conversational character interfaces research has produced widely divergent results and it has tended to focus on the character's dialog capabilities associated with reasoning. Conversely, the present work attempts to evaluate a conversational character from interaction-related aspects. This paper describes an analysis of a conversational character spoken dialogue system using a discrete event simulator. The simulation model was implemented in the ARENA traditional simulator. Simulation results show that sometimes the response time from the speech conversational character can be too long to user, as well as environment noise is an important aspect of the system to be improved.

Keywords: Virtual Reality (VR), Virtual Humans, Conversational Characters, Natural User Interface (NUI), Speech Interface, Conversational Interaction, ARENA.

1 Introduction

Virtual reality technology has become a very popular technology, which embodies the newest research achievements in the fields of computer technology, computer graphics, sensor technology, ergonomics and human-computer interaction theory. Virtual reality and interactive technology have attracted a great deal of attention and much active research is currently being carried out in an effort to investigate their possible benefits in several areas. It is not a secret: virtual reality technology is an important technology to be paid attention, which will bring huge impact to our life and work [1–4].

New developments in the fields of speech recognition, natural language processing, and computer graphics have given rise to the emergence of more sophisticated computer interfaces with multimodal interaction [5].

On one hand, the evolution in speech technologies allow the development of new systems with several purposes, including voice search, personal digital assistant, gaming, living room interaction systems, and in-vehicle infotainment systems as the most popular applications in this category [6].

On the other hand, embodied conversational character emerged as an specific type of multimodal interface, where the system is represented as a person conveying information to human users via multiple modalities such as voice and hand gestures, and the internal representation is modality-independent, both propositional and non-propositional. Embodied conversational character answers questions and performs tasks through interaction in natural language-style dialogs with users contrasting the traditional view of computers. Many people believe that such interfaces have great potential to be beneficial in human-computer interaction for a number of reasons. Conversational character could act as smart assistants, much like travel agents or investment advisors [7–10].

Conversational character applies rich style of communication that characterizes a human conversation. Researchers have built embodied multimodal interfaces that add dialogue and discourse knowledge to produce more natural conversational characters. For example, *Peedy the parrot* is an embodied character that allows users to verbally command it to play different music tracks. While most research on embodied conversational characters has concentrated on the graphical representation and conversational capabilities of the virtual character, others investigated the question of whether auditory embodiment can provide cues that influence user behaviors and ultimately affects the learning performance of users interacting with a virtual character to learn about biology, mathematics, geography, literature, etc. [11–14].

Over last decade, virtual character has been increasingly used for educational purposes. The rationale behind this emerging trend is the belief that information technology can be utilized as a powerful means to assist learners with the acquisition of general knowledge, literacy, narrative competence, social skills like teamwork and negotiation capabilities, logical and spatial reasoning, eye-hand coordination and fine motor control, among others [15–17].

Conversational character must allow the user (speaker) to watch for feedback and turn requests, while the character (listener) can send these at any time through various modalities. The interface should be flexible enough to track these different threads of communication in the appropriate way to each thread. Different threads have different response time requirements; some, such as feedback and interruption occur on a sub-second timescale [18].

It is evident that human-computer communication is extremely complex just as human-human communication. Its analysis implies to turn to several disciplines like as psychology, sociology, biology, among others for foundation of computational model. These disciplines provide very precious qualitative and quantitative information that are indispensable to consider. Moreover, the evaluation of single modalities from conversational character often can be useful for system performance analysis. In this case, simulation models are a powerful tool for both understand the system behavior and to detect possible bottlenecks [19–22].

Thus, as the conversational interaction is a vital aspect for an embodied conversational character, in the present work we present a simulation approach as a real-time modeling for the speech modality.

2 Conversational Character System

The spoken dialogue subsystem of a semantic conversational character is the system simulated. The *character* feature means that system's interface has a visual embodiment and the *conversational* feature means that the system simulates a conversational skill like a human being so the system shows listening and speaking ability. Finally, the conversational character is *semantic* because it is able to perform reasoning through a query to a *Dbpedia*'s query module for users' answers [23, 24].

More specifically, the implemented approach is a question-answering virtual character driving responses to user's inquiries (See fig. 1). The system, named CAVE-VOX, is thought to be used in inquiry applications through the analysis of the user's keyword natural language utterance and the generation of the appropriate response; previous works describe the whole system [25, 26].

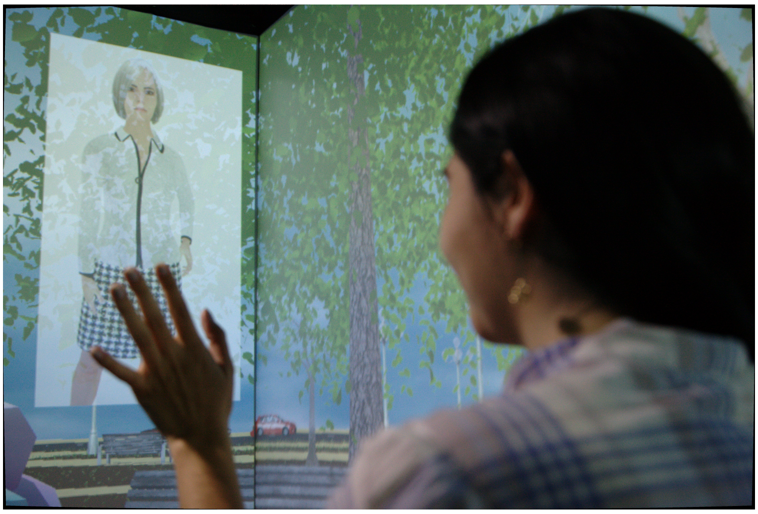

Fig. 1. CAVE-VOX system.

System's human-computer interaction involves real-time synchronization of several aspects like visualization, speech recognition, keywords detection, speech synthesis, among others. Thus a dialogue system between user and virtual character that exploits knowledge provided by structured data (ontologies) in order to help users in an specific information search was built. Then, this question-answer character has a collection of responses relevant to a particular topic. The driven query must be an specific query about a previous chosen topic. The character plays an appropriate role according with the selected topic and answers to user's queries through data stored in a *Dbpedia* data base.

2.1 Speech system

In this work, the spoken dialogue subsystem of a semantic conversational character was simulated. As the real-time interaction is important for this analysis the conversational module was only considered, this system's modality involves simulating inherent human skills like as listening and talks.

The interaction between the character and the user is performed in real-time through user's voice (by speech to text conversion) and virtual character's voice (by text to speech translation), adding simple gestures and expressions, and lip synchronization.

In order to make the evaluation and considering the original system, the main involved components are:

- *Speech recognition (listening)*: while interacting, a human user talks to the system, which transforms the user's speech to a textual representation by using an **Automatic Speech Recognition**(ASR) module. The module takes as input the user's speech utterance that comes from microphone and gives a resultant text from parsing the word string produced by speech recognition and forms an internal semantic representation based in keywords contained into a grammar.
  Each word said by the user is analyzed for the recognizer and compared through a phonetic transcriber. All the picked up words by the microphone are contrasted with words defined by rules in a grammar. According with the grammar, possible sentences are analyzed and compared with the received sentence, then every word gets a confidence value. The ASR module takes this confidence value to determine if a word is accepted or rejected.
  If the word is accepted, then it is stored as a valid keyword. If the word is rejected, then the system must inform about the failure to user. From previous analysis with users, it is known that the rejection rate is approximately 50%.
  When all the words are accepted it allows for a query as a text string consisting of keywords.
  Because the process performed by the ASR is completely computational and is performed in parallel, it is not possible for the user to know the time required to execute this operation and it is usually represented in nanoseconds. Therefore, for the simulations performed in this work, we consider that the ASR processing time is 0 seconds.

- *Reasoning (thinking)*: this module generates a response based on the input, the current state of the conversation and the dialog history. For this, it extracts the meaning of the utterance from keywords, manages the dialog flow and produces the appropriate actions for the target domain in on-line *Dbpedia*.

This approach gives complete control over the virtual persona's knowledge and expressions to the scriptwriter who creates the responses. It allows the writer to specify the character of the virtual persona, what information it can deliver and the form of that delivery. When an interactor comes up to the conversational character and asks it a question, the system driving the character analyzes the interactor's question and selects the appropriate response from *Dbpedia* collection.

Finding an answer takes between 10 and 120 seconds according with the user's query. If this process faults, then the system notifies to the user about the fault. From previous analysis, it is known that the fault rate is approximately 10%.

- *Speech synthesis (talking)*: It is a **Text-To-Speech** (TTS) module carrying out the generation of the synthetic output voice from the text that comes as a response from the Reasoning Module. This module is performed on three main situations: if one word is rejected, then the conversational character says "*Excuse me, I didn't hear you very well*" which takes about 3 seconds, if the *Dbpedia* query faults, then character says "*Sorry. It's not possible to retrieval the elements you requested*" which takes about 7 seconds, and if the conversational character has an answer from *Dbpedia*, then the delay is between 40 and 300 seconds according with its length.

3 Modeling and Simulation with ARENA

Modeling and simulation provide the basis for efficient solving of various problems related to the operation of complex systems like analysis, optimization and management, industry problems (like mining process, etc.). Simulation is considered to be one of the most effective technologies for the analysis and planning of logistics systems [27].

ARENA simulation software, is a general propose simulation tool enabling the construction of models over a series of modules or basic components organized hierarchically. The ARENA simulation software has high level of modeling supporting graphical design. It also includes a lower level of modeling including specific details as arrival times, service time, scheduling of processes, etc. [28].

A model is developed using modules that are part of the basic processes. In ARENA, modules are the flowchart and data objects that define the process to be simulated. All information required to simulate a process is stored in modules. The dynamics associated with the processes can be viewed as nodes in a network by which entities circulate causing a change in the system state. The entities with attributes and variables compete for the services provided by the resources. Entities are items (like trucks, mineral, etc.) that are being served or produced [29].

Figure 2 shows a simple model build with ARENA. The CREATE module is the starting point for entities in a simulation model. Entities are created using a schedule or based on a time between arrivals. Entities then leave the module to begin processing through the system. The entity type is specified in this module. The PROCESS modules are intended as the main processing method in the simulation. They include the resource by which entities compete. A resource retained by an entity must be released at some point in the model. Otherwise, a deadlock can occur. The DECIDE module allows for decision-making processes in the system. Finally, the DISPOSE modules are ending point for entities in a simulation model [30].

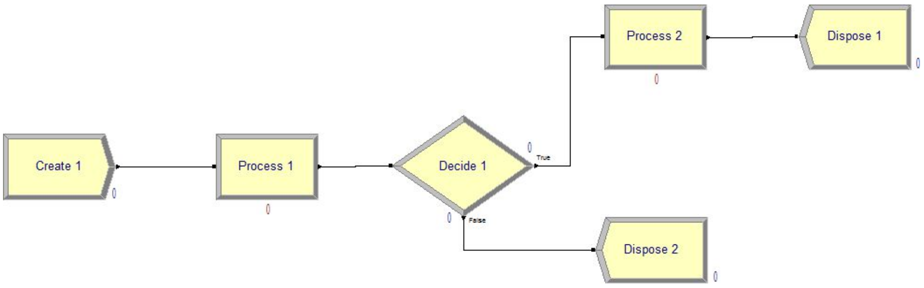

Fig. 2. Basic modules used in ARENA.

After the model is built, we can run simulations to obtain different metrics and statistics like resources utilization, waiting time, etc.

4 Conversational Character Modeling

The simulation model covers a subset of a conversational character. ARENA v.10 was used for modeling the real-time simulation of the spoken dialogue system. As ARENA is an entity-driven application, only the tasks that directly impact the entities are modeled. In the model, an entity represents a query made by the user.

Our resulting model contains the three basic components of a conversational system:

- *Speech recognition (listening)*: this system's stage modeling uses a CREATE module which creates entities (queries based on a time between arrivals), a PROCESS module intended to represent the process performed by ASR resource (because it is not possible for the user to know the time required to execute this operation then the simulated time-delay is zero), and a DECIDE module representing the process which decide as to whether or not something heard by the recognizer is a valid word (according with this problem the success rate is approximately 50%).

- *Reasoning (thinking)*: this component is modeled with a PROCESS module representing keyword searching in the *Dbpedia* resource (with a uniform delay between 10 and 120 seconds), and a DECIDE module determining as to whether or not a keyword is in *Dbpedia* (according with this problem the success rate is approximately 90%).

- *Speech synthesis (talking)*: it is represented by means of three different PROCESS modules according with the three main established situations (3 seconds, 7 seconds and a uniform delay between 40 and 300 seconds respectively). All of these processes use TTS resource and all of them complete at DISPOSE module.

Considering the system's components described before, our simulation model is showed in figure 3.

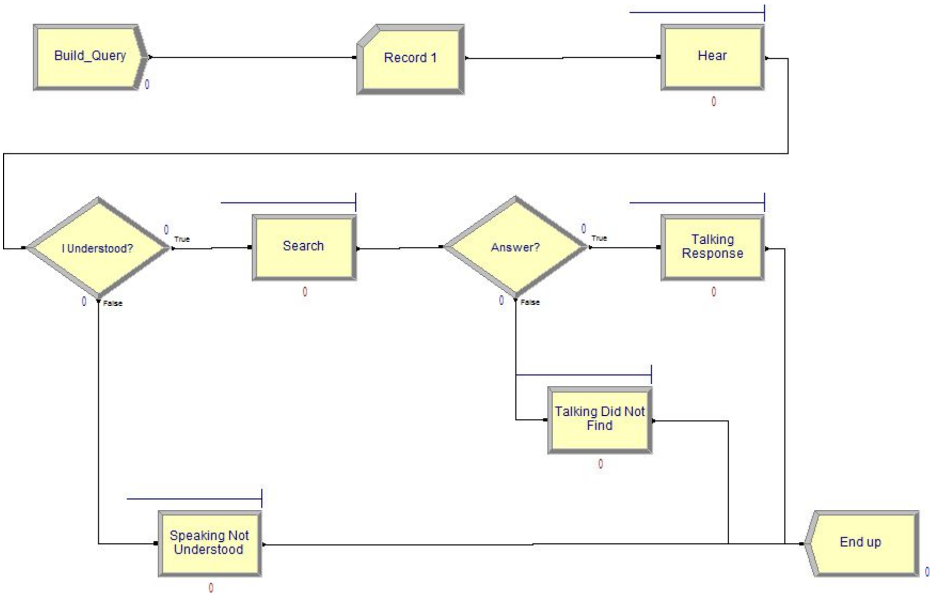

Fig. 3. Conversational Character modeling in ARENA.

5 System Evaluation

Once the simulation model of the conversational character has been built and verified, it can be used to analyze the system performance. After the simulation's running, the ARENA simulator recorded several data related to the model parameters. ARENA reports include category by replication, entity times, entity number, queue times, resource usage, among others.

In this case, 20 replications were performed. From every replication, information about entities times and resources usage was collected: entity parameter represents the query from user and resources parameters are ASR, TTS and *Dbpedia* tools.

The simulation model generates outputs including the performance measure used in the users' experiments considering the process realized by the conversational character from the user's utterance to the appropriate response synthesized by the subsystem for a particular situation.

Query is a the unique entity in the simulated system. In this system the query analysis is vital because a query time allows to learn about response times of the system. The analyzed time for the entities (query) is from the query creation to the query output. An average time from all repeats was considered to obtain minimum, maximum and average times of entity. Table 1 shows the resulting values in seconds.

	Minimum	Maximum	Average
Value	3,0	359,2	103,4

Table 1. Times entity: Query.

From these values, it is concluded that the processing time necessary to attend the demand of a user is 103 seconds, meaning near 2 minutes in average, and the maximum time is 359 seconds meaning near 6 minutes. Minimum value represents the case when speech recognition fails.

In analysis of resources, resource usage was considered. ASR and TTS resources are used in each query only once, then their constant values are not important for system's performance analysis. This is different from *Dbpedia* resource where its analysis requires to be considered. According with ARENA outputs, *Dbpedia* usage is about 30% as a minimum, 80% as a maximum and 42,3% in average. It is important to emphasize that if *Dbpedia* resource is not used this is due to speech recognition failure.

6 Conclusions and Discussion

Latest research had been oriented to create embodied computer-animated characters that produce accurate auditory and visible speech, as well as realistic facial expressions, emotions, and gestures. The invention of such characters has a tremendous potential to benefit virtually all individuals in natural user interface. Conversational interaction is an important aspect to enable and improve the human-computer interaction.

This paper involved the evaluation of the spoken dialogue system of a conversational character with a discrete event simulator ARENA.

We described character's speech system and its simulation model that was developed to study conversational interaction performance. We introduced the general and most important issues one has to take care of when starting to simulate with simulator ARENA. We discussed the modules and parameters in detail with the main idea to obtain quantitative results from simulation.

At the moment, analysis is limited to entities and resources parameters. Even though, the analyses does not cover all the aspects of the user query, we believe it has several elements to improve. For example, from entity point of view the simulation results show that sometimes the extensiveness of the speech from the conversational character can be too long to users. A more exhaustive study is need for determining if a long answer is produced by reasoning or something else. For the analysis of resources, *Dbpedia* usage is not ideal because sometimes speech recognition fails due to many factors influence in a correct recognition of speech as: speaker's clarity, microphone's quality, user's accent, environment noise, etc. Several of these aspects can be improved by upgrading microphone's quality.

In this work, we focused on human-computer conversational interaction in real-time and how to do it more natural and intuitive to users allowing proficient user-interactivity in real-time, meaningful feedback and learning through an interface.

References

1. XinXing Tang, editor. *Virtual Reality - Human Computer Interaction.* InTech, 2012.
2. M. Chan. *Virtual Reality: Representations in Contemporary Media.* Bloomsbury Publishing, 2014.
3. T. Parisi. *Learning Virtual Reality: Developing Immersive Experiences and Applications for Desktop, Web, and Mobile.* O'Reilly Media, Incorporated, 2015.
4. Kurt Squire. Changing the game: What happens when video games enter the classroom? *Innovate: Journal of Online Education*, 1(6), August 2005.
5. Cecilia Sik Lanyi, editor. *The Thousand Faces of Virtual Reality.* InTech, 2014.
6. Dong Yu and Li Deng. *Automatic Speech Recognition: A Deep Learning Approach.* Springer Publishing Company, Incorporated, 2014.
7. J. Cassell, T. Bickmore, L. Campbell, H. Vilhjálmsson, and H. Yan. Embodied conversational agents. chapter Human Conversation As a System Framework: Designing Embodied Conversational Agents, pages 29–63. MIT Press, Cambridge, MA, USA, 2000.
8. M. Mancini. *Multimodal Distinctive Behavior for Expressive Embodied Conversational Agents.* Universal Publishers, 2008.
9. Q. Chen, P. Torroni, S. Villata, J. Hsu, and A. Omicini. *PRIMA 2015: Principles and Practice of Multi-Agent Systems: 18th International Conference, Bertinoro, Italy, October 26-30, 2015, Proceedings.* Lecture Notes in Computer Science. Springer International Publishing, 2015.
10. B. Endrass. *Cultural Diversity for Virtual Characters: Investigating Behavioral Aspects across Cultures.* EBL-Schweitzer. Springer Fachmedien Wiesbaden, 2014.
11. Niels Ole Bernsen and Laila Dybkjr. *Multimodal Usability.* Springer Publishing Company, Incorporated, 1st edition, 2009.
12. Gene Ball, Dan Ling, David Kurlander, John Miller, David Pugh, Tim Skelly, Andy Stankosky, David Thiel, Maarten V Dantzich, and Trace Wax. Life-like computer characters: The persona project at microsoft research. *Software agents*, pages 191–222, 1997.

13. Sharon Oviatt, Courtney Darves, and Rachel Coulston. Toward adaptive conversational interfaces: Modeling speech convergence with animated personas. *ACM Trans. Comput.-Hum. Interact.*, 11(3):300–328, September 2004.
14. R.J.S. Sloan. *Virtual Character Design for Games and Interactive Media.* CRC Press, 2015.
15. Andrea Corradini, Klaus Robering, and Manish Mehta. *Conversational characters that support interactive play and learning for children.* INTECH Open Access Publisher, 2009.
16. David Griol, José M Molina, Zoraida Callejas, and Ramón López-Cózar. La plataforma educagent: agentes conversacionales inteligentes y entornos virtuales aplicados a la docencia. 2011.
17. Marissa Milne, Martin Luerssen, Trent Lewis, Richard Leibbrandt, and David Powers. Embodied conversational agents for education in autism. *A comprehensive Book on Autism Spectrum Disorders*, page 387, 2011.
18. Dominic W Massaro, Ying Liu, Trevor H Chen, and Charles Perfetti. A multilingual embodied conversational agent for tutoring speech and language learning. In *INTERSPEECH*, 2006.
19. J.O. Turner, M. Nixon, U. Bernardet, and S. DiPaola. *Integrating Cognitive Architectures into Virtual Character Design.* Advances in Computational Intelligence and Robotics. IGI Global, 2016.
20. Y. Zhang, P. Zhang, and D.F. Galletta. *Human-computer Interaction and Management Information Systems: Foundations.* Taylor & Francis, 2015.
21. Benjamin Weiss, Ina Wechsung, Christine Kühnel, and Sebastian Möller. Evaluating embodied conversational agents in multimodal interfaces. *Computational Cognitive Science*, 1(1):1, 2015.
22. Lee W Lacy. *Interchanging Discrete event simulation Process Interaction Models using the Web Ontology Language-OWL.* PhD thesis, University of Central Florida Orlando, Florida, 2006.
23. Yu-Lin Chu and Tsai-Yen Li. *Realizing semantic virtual environments with ontology and pluggable procedures.* INTECH Open Access Publisher, 2012.
24. Mohamed Morsey, Jens Lehmann, Sören Auer, Claus Stadler, and Sebastian Hellmann. DBpedia and the Live Extraction of Structured Data from Wikipedia. *Program: electronic library and information systems*, 46:27, 2012.
25. Y. Alvarado, N. Moyano, D. Quiroga, J. Fernández, and R. Guerrero. *Augmented Virtual Realities for Social Developments. Experiences between Europe and Latin America*, chapter A Virtual Reality Computing Platform for Real Time 3D Visualization, pages 214–231. Universidad de Belgrano, 2014.
26. N. Jofré, G. Rodríguez, Y. Alvarado, J. Fernández, and R. Guerrero. Virtual humans: Conversational characters for a cave-like environment. In *XX Congreso argentino de ciencias de la computación*, pages 937–946. Universidad Nacional de La Matanza, Octubre 2014.
27. Mustafa Rawat and Steven Vaccaro. Implementing a distributed logistics simulation using arena and hla. In *Proceedings of the 2009 Spring Simulation Multiconference*, page 62. Society for Computer Simulation International, 2009.
28. T. Altiok and B. Melamed. *Simulation Modeling and Analysis with ARENA.* Elsevier Science, 2010.
29. M.D. Rossetti. *Simulation Modeling and Arena.* Wiley Series in Modeling and Simulation Series. Wiley, 2015.
30. W.D. Kelton, R.P. Sadowski, and N.B. Swets. *Simulation with Arena.* McGraw-Hill Education, 2015.