A Virtual Animated Commentator Architecture for Cybersecurity Competitions

8

Ruth Agada, Jie Yan, and Weifeng Xu

Abstract

Cybersecurity competitions are exciting for the game participants; however, the excitement and educational value do not necessarily transfer to audiences because audiences may not be experts in the field. To improve the audiences' comprehension and engagement levels at these events, we have proposed a virtual commentator architecture for cybersecurity competitions. Based on the architecture, we have developed a virtual animated agent that serves as a commentator in cybersecurity competition. This virtual commentator can interact with audiences with facial expressions and the corresponding hand gestures. The commentator can provide several types of feedback including causal, congratulatory, deleterious, assistive, background, and motivational responses. In addition, when producing speech, the lips, tongue, and jaw provide visual cues that complement auditory cues. The virtual commentator is flexible enough to be employed in the Collegiate Cyber Defense Competitions environment. Our preliminary results demonstrate the architecture can generate phonemes with timestamps and behavioral tags. These timestamps and tags provide solid building blocks for implementing desired responsive behaviors.

Keywords

Virtual agent · Software architecture · Cybersecurity · Education · Animation

R. Agada (✉) · J. Yan · W. Xu
Department of Computer Science, Bowie State University, Bowie, MD, USA
e-mail: ragada@bowiestate.edu; jyan@bowiestate.edu; wxu@bowiestate.edu

8.1 Introduction

Cybersecurity is a field that is continually garnering much interest as it is now pervasive in the everyday lives of people. To educate the public and train prospective security specialists, the academic community and industry has been responding by developing new programs in information assurance and devising creative ways to attract and train the next generation of cybersecurity professionals [1–3]. One such means has been to create a cybersecurity competitions [4]. These competitions serve to educate its participants and spectators alike. However, the number of student participants engaged in these competitions has still been relatively small. One of the reasons for this lack of interest may be attributed to the fact that, to date, the competitions have been beneficial mainly to anticipants, such as student teams, cybersecurity experts, and administrators and judges of the games. However, for an audience of the sport, the excitement and educational value do not necessarily transfer. This may be because there are audiences who are not experts in the field, and the information and visualization tools are too high level for them to understand.

To help the audiences to comprehend cybersecurity competitions and encourage their engagements at these events, many educators use video game format [5] to educate their participants. Researchers have observed that audiences become engaged in the activity and gain educational content vicariously [2, 6, 7].

Considering the effectiveness of applying video game in education, this paper presents a system to build a game-like virtual commentator that aims to help non-expert audiences comprehend the concepts and engage audiences in cybersecurity-related events, and therefore, promote cybersecurity education among non-expert audiences. Specifically, the commentator is able to: (1) perform various human-like behaviors which range from various valenced facial

S. Latifi (ed.), *Information Technology – New Generations*, Advances in Intelligent Systems and Computing 738,
https://doi.org/10.1007/978-3-319-77028-4_8

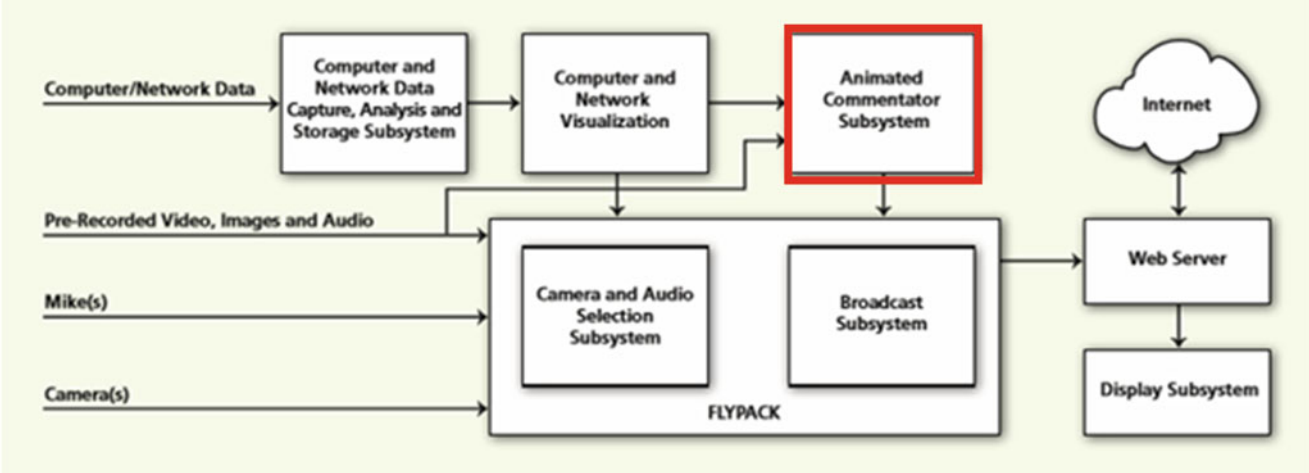

Fig. 8.1 Logical view of the LUCID visualization system

expressions to gestures, gaze and emotions conveyed in speech, (2) interacting with audiences, and providing several types of feedback including causal, congratulatory, deleterious, assistive, background, and motivational responses, and (3) when producing speech, the lips, tongue, and jaw provide visual cues that complement auditory cues. In addition, the virtual commentator is flexible enough to be employed in the Collegiate Cyber Defense Competitions environment.

NSF mainly supports this research "LUCID: A Spectator Targeted Visualization System to Broaden Participation at Cyber Defense Competitions". The proposed virtual commentator is built as part of the LUCID framework shown in Fig. 8.1. The goal of this project was to stimulate interest in cybersecurity competitions through the development of a visualization and broadcast system targeted to enhancing learning and understanding among spectators. The LUCID framework is comprised of five basic subsystems: a computer and Networking data capture, analysis, and storage subsystem, a computer and networking visualization system, a camera and audio selection subsystem, a broadcast subsystem, and display subsystem. Each system collects their respective data and in concert with one another relaying the necessary information in some fashion to the spectator [2].

The animated commentator subsystem, highlighted by the red box in Fig. 8.2, takes video, images, and audio information, as well as computer and network visualization information, to generate semantic tags, that control various behaviors from facial expressions to gestural motion of animated agent to make it believable, personable and emotional. These behaviors are broadcasted and displayed on the camera and audio subsystem.

The rest of this paper is organized as follows: Sect. 8.2 describe the architecture of the commentator subsystem. Section 8.3 demonstrates some preliminary development results.

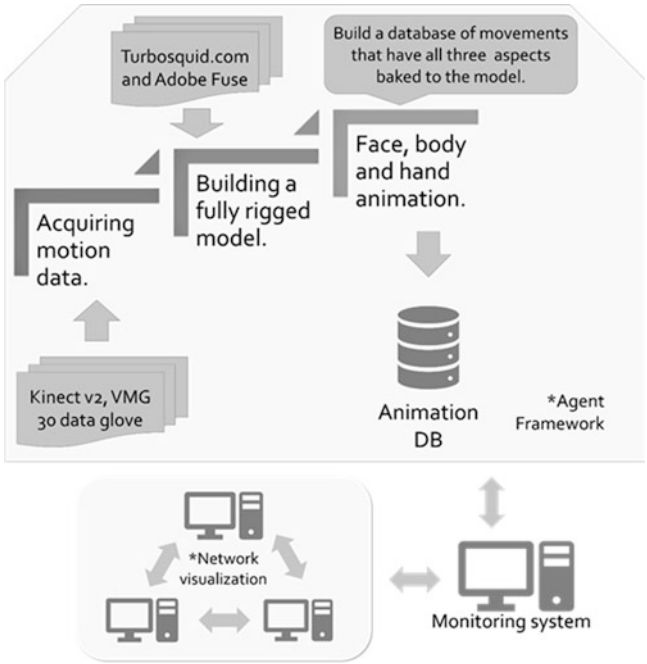

Fig. 8.2 Animated commentator subsystem architecture

Section 8.4 reviews the related work. Section 8.5 concludes the paper.

8.2 The Commentator Subsystem

As stated earlier, the commentator subsystem in the LUCID framework aims to engage the spectators in a similar fashion [8] as with video games. The commentator subsystem focuses on the visual and auditory component in terms of a virtual environment.

8.2.1 Virtual Commentary Architecture

The architecture of the animated commentator subsystem is shown in Fig. 8.2.

The core of the architecture consists of three components:

- An integrated hardware and software interface for acquiring motion data from a variety of hardware device, including Kinect v1 camera and VMG 30 data glove.
- A software system to build a fully rigged behavioral model for the commentary based on motion data.
- An animation software to combine face, body and hand tags for the agent's behavioral models.

For modeling different affective states exhibited by most engaging sports casters, we develop a system to generate images in the 3-D scene in two steps. Firstly, in the modeling step, the system produces a precise description of the agent, in terms of graphics primitives. Then it acquires the vertex data required for the primitives from 3-D modeling software that can generate vertex data. The data provided for the development of the animated agent contain over one thousand points in 3-D space. Secondly, in the rendering step, the vertex data also serve to draw the model to the screen.

For the system to simulate a smooth transition in the model's facial expression, as well as different full body gestures, our method generates a sequence of vertices given that only known vertices are the start and end vertices. By studying and annotating the video clips we collected, we can model realistic animation patterns. With the help of third-party applications, the animation is natural-looking and includes head and face movements, combined with the movements of facial components (eyes, eyebrows, nose, cheeks, etc.) and full body movements (arms, hands, fingers, torso, legs, etc.).

8.2.2 Motion Generation

For the agent to simulate human commentator behavior, initially, we used video footage of various contact sports commentators. Sample footage of commentator came from NFL roundtable discussions, super bowl commentators, and Apollo Robbins' TED talks. Each video file is roughly 9 min in length. These contain commentator motion data that can be applied to the animated agent. The videos used in this project were sourced from YouTube.com. To further analyze behaviors of commentators from non-contact sports, especially from cybersecurity events, we collected data from the Maryland Cyber Challenge. In 2014, we attended one such event in which we interviewed (and recorded) one of the sponsors of the competition. The interview lasted 30 min, and from that, we created frame-by-frame shots to aid in the analysis of facial and body gestures. Again in 2015, we performed a similar data collection, which provided us with data 1 h and 30 min in duration.

We observed several motions/gestures across all the collected data. To annotate the footage, we used a tool called ELAN [9]. With ELAN [9], we created three tiers to represent the different commentator behaviors we analyzed. Further to that we added another tier to represent the dialog phase of the interaction between the interviewer, spectators, and event sponsors. Figure 8.3 shows the interface of ELAN and the different tiers which allow the user to analyze different media corpora by adding descriptive tags that are either be time-aligned to the media or it can refer to other existing annotations.

From the analyzed video clips, we observed typical animation pattern for a commentator with individual personalities. We represent this animation pattern by using tagged meta-animation sequence. Based on behaviors of interest in the expression and gestural behaviors, we developed tagged animations sequences that correspond to the tier of behavior. To capture the requisite gestures and facial expressions used by sports commentators several hardware devices and software applications. Microsoft's Kinect v2 captured face and body motion data, while the virtual motion glove—VMG30—captured relevant hand motion. To process the data stream from both the Kinect v2 and VMG30, we used the Brekel and VMG SDK to map the input motion data to the animated model.

8.2.3 Speech Generation

The animated agent is required to give commentary on the cybersecurity competition, and to give this commentary while maintaining human-like expressions, hand gestures, and movements. The facial expression demonstrated by the agent is dependent on sentences being narrated by the agent. Having this work seamlessly for the different scenarios that may arise based on the activities in the competition involves two parts. Firstly, a database that provides the appropriate triggers for the agent to speak based on different conditions in the form of text, and secondly, a system that takes this text and converts it into speech while using the appropriate facial expressions.

The database used in this system is the MySQL database. For our system, the events happening during the competition, mainly the services that are up or down (up indicating this it is currently still running, and down indicating that the service has been compromised and shut down), cause data to be stored in the MySQL database. This data contains keywords that the agent needs to narrate. We created scripts that query

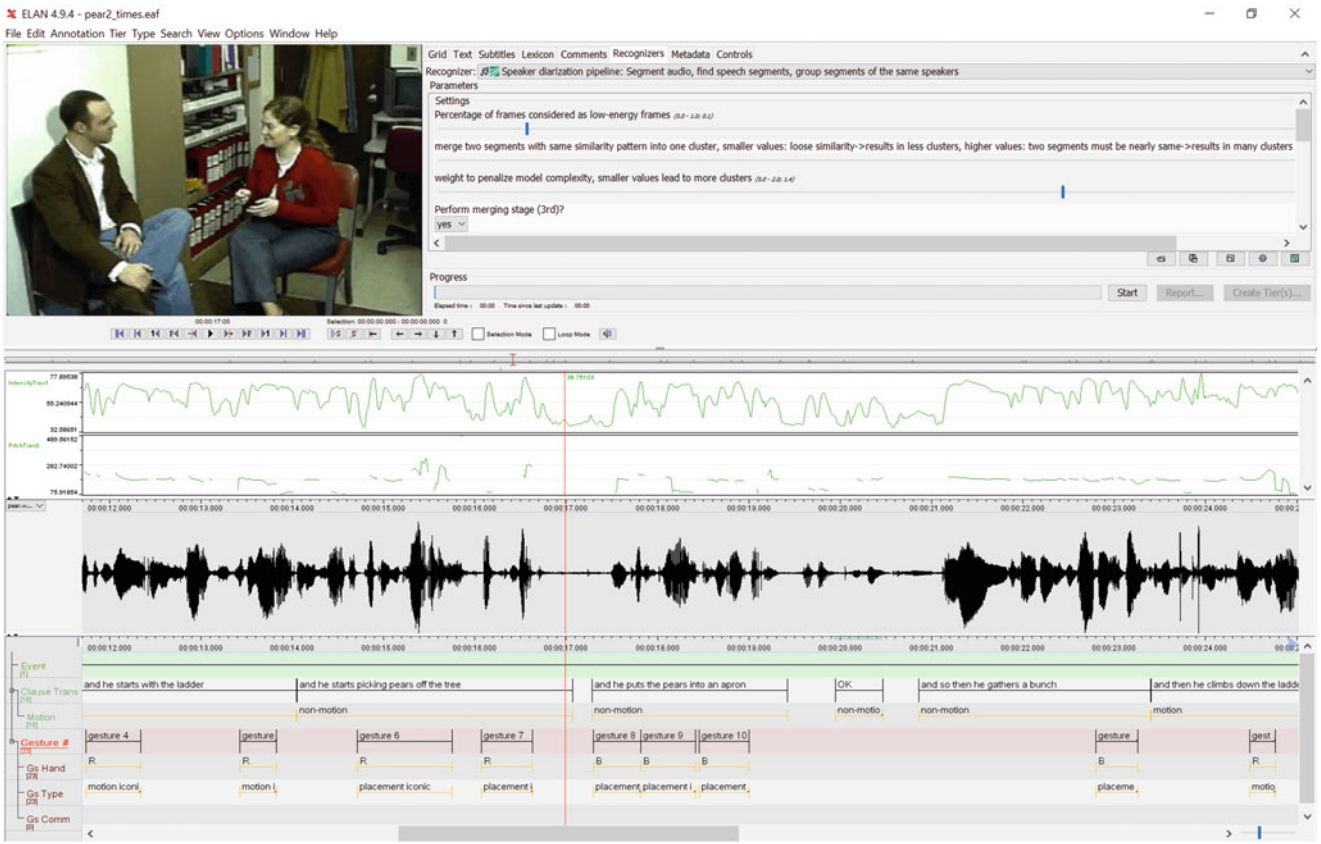

Fig. 8.3 Annotated video clip using ELAN and the associated annotation tiers

this database whenever the virtual agent needs to announce information to the spectators.

The second part of the system is the TTS. For this part of our system, we used several third-party systems to provide speech functionality and to ensure that the text spoken by the agent also creates appropriate lip movement and facial expressions to generate a full and realistic virtual commentator. We use two separate software for the text-to-speech because of the different advantages they provide. MaryTTS [10, 11] is an open-source, multilingual Text-to-Speech Synthesis platform written in Java. Being open-source means that it is available for free online. We chose this because of its availability, user-friendliness, and ease of use. The result of the query, i.e., the text, is inputted into this software to generate the necessary phonemes contained in each word. A phoneme is the basic unit of sound in any word in a specified language. The phoneme is used to determine the way the agent's lips move. Phonemes affect how words are pronounced, things like accent, the shape of the mouth, language all affect the way the mouth and lips move. This phoneme list is generated and then used to facilitate the mouth shapes of the agent when pronouncing each word. The second software is CereVoice [12, 13], which is another TTS software. This is also open source and available through different license tiers. We use CereVoice because it produces a much more natural sounding synthesized speech than MaryTTS [10, 11]. It also takes in the same text as input and is responsible for producing the final audio that will be heard by the spectators. To tie the model's speech capabilities to its motion, we use the Lip Sync pro system [14] to provide a window for synchronizing phonemes generated, emotions and gestures to dialogue in an audio file, and a component for playing back these dialogue clips on a character using totally customizable poses.

8.3 Preliminary Results

Our preliminary results include: (1) developed expressions/gestures of an animated virtual commentator and (2) generated a marked-up animation sequence.

8.3.1 Generating Expressions and Gestures

Figure 8.4 shows some of the expressions and gestures of an animated virtual commentator we have developed to engage spectators in the Collegiate Cyber Defense Compe-

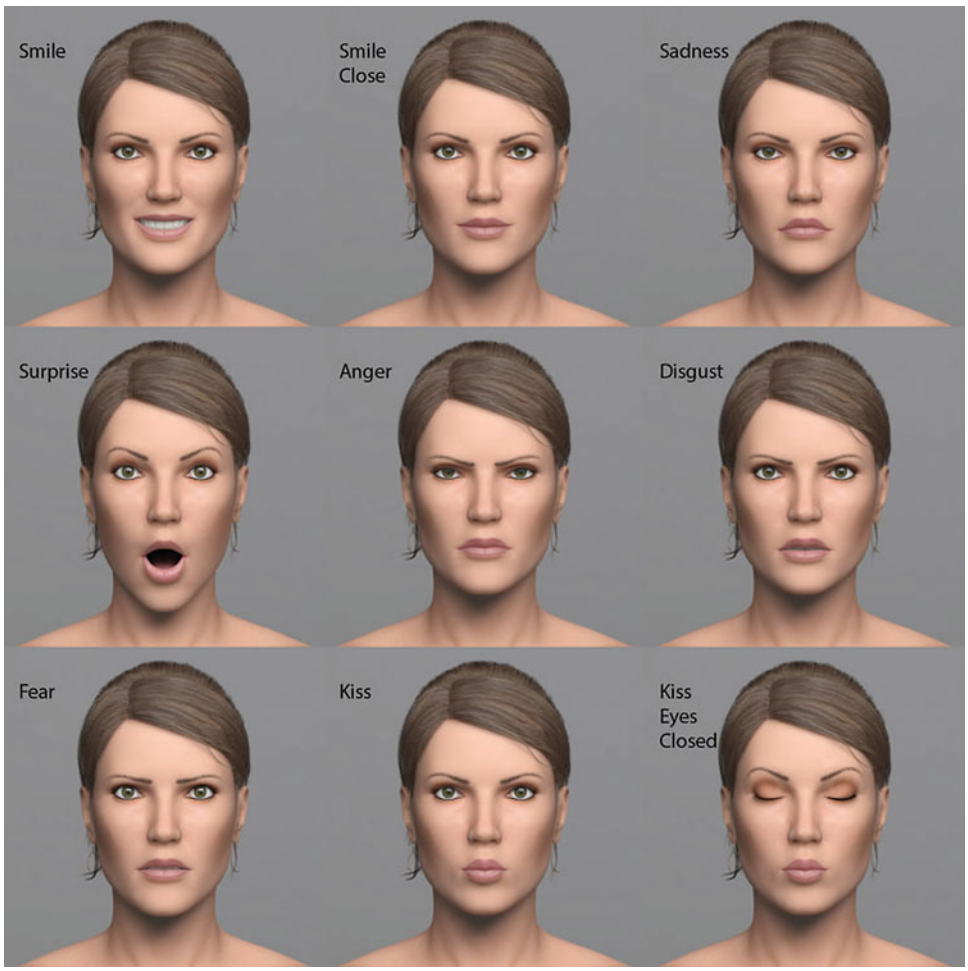

Fig. 8.4 Diverse expression of an emotional embodied conversational agent. Top row: smile with teeth, smile closed mouth, and sadness. Middle row: surprise, anger, and disgust. Bottom row: fear, pursed lips and pursed lips with closed eyes

titions (CCDCs) environment. By recording and observing the interactional behaviors of different professional human commentator working with various groups of spectators in a set of CCDCs, we have identified specific coordinated verbal and nonverbal signals exchanged during the class session. The nonverbal cues include head nods, the direction of eye gaze (including eye contact), facial expressions, posture shifts and hand and arm gestures. Once these have been codified and analyzed to determine their temporal organization and contexts of occurrence, we conducted several experiments and noted the common behaviors exhibited by the observed commentators and mapped those behaviors unto the animated agent.

The current interface, while independent of other external systems, as seen in Fig. 8.5, contains the current virtual environment for one of the virtual commentators created for the system. The current scene is a 3D mock-up of the bridge of the starship Enterprise from the popular science fiction show Star Trek™: The Next Generation. On initialization of the system, the agent checks if it has established a connection to the MySQL database that houses the data collected from the network monitoring system. If there is any connection issue, the agent reports that to the spectator in randomly selected over emphasized gestures with matching speech generation as seen in Fig. 8.6.

8.3.2 Generate a Marked-up Animation Sequence

To generate a marked-up animation sequence or to add additional markup to a sequence, we need to insert a series of animation tags to the phoneme.

Here is an example to describe how it works as seen in Table 8.1. Suppose the input text string is the prompt "didn't know you know so much," with associated phoneme string and animation sequence and, that a meta animation sequence will be imposed on this animation sequence automatically.

Using for example, the meta animation sequence example: 0.2 s—<FET = SMILE>—0.2 s—<EYEB = (3000, B)>—0.2 s—<HGT = (−2,0,0)>—0.3 s—<HGT = (2,0,0)> <FET = RAISEEYEBROW>—0.3 s—<HGT = (4,0,2)>—0.3 s—<FET = SMILE>. The smile target is inserted on a phoneme boundary close to the 0.2 s into the animation as

Fig. 8.5 System interface with a virtual agent in an idle animation

Fig. 8.6 Top: Model behavior for successful MySQL connection by showing an exaggerated thumbs up gesture and returning to the idle animation. Bottom: Model behavior for failed MySQL connection by showing disappointment by shrugging its shoulders

Table 8.1 Phoneme generation with time stamps and no behavioral tags

Phoneme	Duration	Word boundary	Tag<s>
D	0.054	1	
I	0.09	0	
d	0.08	0	
&	0.06	0	
n	0.06	0	
n	0.1	1	
oU	0.2	0	
j	0.12	1	
u	0.069	0	
n	0.04	1	
u	0.27	0	
s	0.14	1	
oU	0.14	0	
m	0.109	1	
ˆ	0.23	0	
tS	0.348	0	
. pau	0.4	1	

Table 8.2 Phoneme generation with desired effective behavioral tags

Phoneme	Duration	Word boundary	Tag<s>
D	0.054	1	
I	0.09	0	<FET = SMILE>
d	0.08	0	
&	0.06	0	
n	0.06	0	<EYEB = (3000, B)>
n	0.1	1	<HGT = (−2,0,0)>
oU	0.2	0	
j	0.12	1	
u	0.069	0	
n	0.04	1	<HGT = (2,0,0)>
			<FET = RAISEEYEBROW>
u	0.27	0	
s	0.14	1	
oU	0.14	0	<HGT = (4,0,2)>
m	0.109	1	
ˆ	0.23	0	<FET = SMILE>
tS	0.348	0	
.pau	0.4	1	

specified by the meta-animation sequence but not at exactly 0.2 s as such, allowing semi-random behavior and ensuring temporal variability while ensuring speech synchronized behavior. The animation sequence may alternatively have specified that the tag be inserted only at a word boundary, probabilistically or in other ways.

Tags in the meta-animation sequence are inserted in order, and the movements from all channels are finally combined to generate a final animation sequence. For the previous example the result may look like Table 8.2.

8.4 Related Work

Cybersecurity competitions are becoming increasingly prominent parts of the landscape of cybersecurity education. Most notably is the National Collegiate Cyber Defense Competition (NCCDC) with its precursor regional events [6] and the UCSB International Capture the Flag Competition (iCTF) [7]. These are both large national and international competitions with scores of institutions and hundreds of students [4]. There are several other competitions held at regional levels [4] and those organized by other faculty and security specialists [7, 15–21].

As O'Leary noted [4], these competitions require the students a certain level of competence in defensive and administrative skills. The teams of participating students work to defend identical networks from a group of designated attackers. The team that successfully keeps most of their services on their network, as well as their network operational are declared winners of the event.

Most of these games favor the war game mode of hands-on exercises for participants. These war games are either organized as capture-the-flag (CTF) [22, 23], king of the hill (KOTH), defend the flag—a variation of KOTH, computer simulations and online programming-level war-games [24, 25]. The issue with these modes is there is no opportunity for the spectator to actively participate in the competition, granted they are prevented from engaging with actual participants of the competition. The lack of a system that involves spectators in the event continues to unintentionally exclude that potential population of security specialists in a field that suffers from reduced engagement from different groups. Hence, we propose the use of an effective pedagogical virtual agent with commentator status that engages the spectators.

8.5 Conclusion

This paper presents a virtual animated commentator architecture for cybersecurity competitions. We have developed a commentator based on the architecture. The commentator is able to generate expressions/gestures and a marked-up animation sequence. Specifically, we have identified a set of verbal and nonverbal signals and cues exchanged during the interactions between spectators and the commentator. To that end, we take advantage of the capabilities the Kinect v2 and VMG 30 glove to model those behaviors and map them to our 3D animated models to better simulate the realistic performance of human commentators. We anticipate that the system will ultimately improve the educational value and excitement for the spectator and broaden interest in

the field of cybersecurity. Our future work is based on two hypotheses in terms of the audience's interaction with the system and the method to educate and engage them. For the spectator, we hypothesize that the excitement and learning outcomes are associated with the ability to extract, visualize and comprehend the details of the game as they unfold will improve [2], thus opening cybersecurity events to a broader group. Secondly, the use of a pedagogically effective animated commentator will aid in the comprehension and engagement at these events.

Acknowledgment This work is mainly supported by grant NSF-DUE 1303424 and partially supported by grant NSF-HBCU-UP 1714261.

References

1. R.S. Cheung, J.P. Cohen, H.Z. Lo, F. Elia, V. Carrillo-Marquez, Effectiveness of cybersecurity competitions, in *Proceedings of the International Conference on Security and Management (SAM)*, (2012), p. 1
2. C. Turner, J. Yan, D. Richards, P.O. Brien, J. Odubiyi, Q. Brown, LUCID: A visualization and broadcast system for cyber defense competitions. ACM Inroads **6**(2), 70–76 (2015)
3. R. Agada, J. Yan, Leveraging automated animated agent commentary to improve sense-making for novice users at cybersecurity competitions. Natl. Cybersecurity Inst. J. **3**(1), 47–55 (2016)
4. M. O'Leary, Small-scale cyber security competitions, in *Proceeding of the 16th Colloquium for Information Systems Education*, (2012)
5. A. Groen, How video games are becoming the next great North American spectator sport, *arstechnica*, (2012), [Online]. https://arstechnica.com/gaming/2012/09/how-video-games-are-becoming-next-great-north-america-spectator-sport/. Accessed 1 Jan 2017
6. History of CCDC, *National collegiate cyber defense competition*. [Online]. http://www.nationalccdc.org/index.php/competition/about-ccdc/history. Accessed 1 Jan 2017
7. A. Conklin, The use of a collegiate cyber defense competition in information security education. in *Proceedings of the 2nd Annual Conference on Information Security Curriculum Development—InfoSecCD'05*, (2005), p. 16
8. G. Cheung, J. Huang, Starcraft from the stands: understanding the game spectator, in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, (2011), pp. 763–772
9. P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, H. Sloetjes, ELAN: a professional framework for multimodality research, in *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation*, (2006)
10. M. Schröder, J. Trouvain, The German text-to-speech synthesis system MARY: a tool for research, development and teaching. Int. J. Speech Technol. **6**, 365–377 (2003)
11. M. Schröder, M. Schröder, Interpolating expressions in unit selection, in *Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interaction*, **2**(2), 718–720 (2007)
12. CereVoice Engine Text-to-Speech SDK, CereProc. [Online]. https://www.cereproc.com/en. Accessed 1 Jan 2017
13. M. Aylett, C. Pidcock, The CereVoice characterful speech synthesiser SDK. AISB **2007**, 174–178 (2007)
14. LipSync Documentation, Rogo Digital, (2016), [Online]. https://lipsync.rogodigital.com/documentation/. Accessed 5 May 2017
15. G.B. White, D. Williams, The Collegiate Cyber Defense Competition, in *9th Colloq. Inf. Syst. Secur. Educ.*, (2005), pp. 26–31
16. G.B. White, D. Ph, I. Assurance, The National Collegiate Cyber Defense, in *10th Colloquium for Information Systems Security Education*, (2006)
17. T. Rosenberg, W.W. Security, C.O. Brien, The growth of the Mid-Atlantic CCDC: public—private partnerships at work, in *Proceedings of the 12th Colloquium for Information Systems Security Education*, (2008), pp. 72–76
18. A. Carlin, D.P. Manson, J. Zhu, Developing the cyber defenders of tomorrow with regional collegiate cyber defense competitions (CCDC). Inf. Syst. Educ. J. **8**(14), 3–10 (2010)
19. A. Cook, R.G. Smith, L. Maglaras, H. Janicke, SCIPS: using experiential learning to raise cyber situational awareness in industrial control system. Int. J. Cyber Warf. Terror. **7**, (2017)
20. B. Hallaq, A. Nicholson, R. Smith, L. Maglaras, H. Janicke, K. Jones, CYRAN: a hybrid cyber range for testing security on ICS/SCADA systems, in *Security Solutions and Applied Cryptography in Smart Grid Communications*, (2017)
21. A. Furfaro, A. Piccolo, D. Sacca, A. Parise, A virtual environment for the enactment of realistic cyber security scenarios, in *Proc. 2016 Int. Conf. Cloud Comput. Technol. Appl. CloudTech 2016*, (2017), pp. 351–358
22. J. Werther, M. Zhivich, T. Leek, N. Zeldovich, Experiences in cyber security education: the MIT Lincoln Laboratory capture-the-flag exercise, in *Proceedings of the 4th Workshop on Cyber Security Experimentation and Test*, (2011)
23. N. Capalbo, T. Reed, M. Arpaia, RTFn: enabling cybersecurity education through a mobile capture the flag client, in *Proceedings of SAM'11*, (2011), pp. 500–506
24. M.E. Whitman, H.J. Mattord, A. Green, Incident response: planning, in *Principles of Incident Response and Disaster Recovery*, 2nd edn. (Cengage Learning, 2013), p. 131
25. S. Jajodia, S. Noel, P. Kalapa, M. Albanese, J. Williams, Cauldron mission-centric cyber situational awareness with defense in depth, in *2011—MILCOM 2011 Military Communications Conference*, (2011), pp. 1339–1344