Eye Tracking and Speech Driven Human-Avatar Emotion-Based Communication

Saverio Cinieri[1], Bill Kapralos[2], Alvaro Uribe-Quevedo[2], and Fabrizio Lamberti[1]

[1]Dipartimento di Automatica e Informatica, Politecnico di Torino, Turin, Italy.
[2]Software and Informatics Research Centre, Ontario Tech University, Oshawa, Canada.

Corresponding Author Email: Bill.Kapralos@uoit.ca

***Abstract*—Feelings, emotions, and empathy play an important role in many daily activities including verbal and non-verbal communication. Their automatic recognition and interpretation is important in a variety of applications requiring communication skills that are difficult to reproduce in computer-simulated environments, including those involving human-avatar interactions. Our recent work has begun investigating the development of intelligent avatars capable of detecting user (human) emotions to allow for realistic human-avatar interactions within medical-based virtual simulations and serious games. In this paper, we present a system that couples eye tracking and dialogue interpretation to allow for intelligent and realistic human-avatar communication. Although formal testing is required, preliminary results are promising, showing the potential of the system.**

Keywords—Emotion, mood, intelligent avatar, eye tracking, human-avatar interaction, virtual reality.

I. Introduction

The technologies of video games, virtual worlds and social networks have become collectively known as immersive technologies because of their ability to engage users of all ages, driving massive investment into technologies to attract, capture and retain our attention [1]. The advent of affordable computational processing power along with the accompanying decrease in the size of electronic components has led to the decreasing cost and increased availability of consumer-level immersive technologies. This has also led to widespread adoption of virtual simulation serious games (i.e., video games whose primary purpose is education, training, advertising, simulation, or education as opposed to entertainment), in recent years [2]. Amongst the recently available technologies are a wide variety of diverse natural user input devices employing facial, hand, and eye tracking that allow for realistic human-computer interactions [3]. For example, eye tracking devices can determine where a user's eyes are focused, and this information can provide details regarding the user's attention, focus, drowsiness, consciousness, or emotions when studying behavior [4].

Although the strongest demand for immersive technologies currently comes from the entertainment industry (e.g., video games), their features have made them suitable for employment in a wider variety of applications and diverse industries such as healthcare, education, military, and real estate [5]. Furthermore, immersive technologies are currently transforming teaching and learning with positive projections of wider adoption by 2025 [6]. With respect to healthcare, immersive technologies (virtual reality, and serious games in particular), are seeing widespread popularity and use in a variety of applications including medical education (see [7]), and physical rehabilitation (see [8,9]), among others. Central to many of these applications are virtual humans or avatars, that is, digital self-representations, enabling individuals to interact with each other within virtual environments [10]. Avatars can represent the user or other non-player characters (NPCs)), which form the "inhabitants" of the virtual (simulated) environment. By adopting personalized avatars, users can interact with each other in virtual worlds and foster cohesive social relations [11]. In addition, interactions with avatars can lead to changes in emotions and cognitive function, and an increase in empathy, emotion regulation, compassion, and connection with others [12]. Avatars are becoming more prevalent not only in virtual environments, but also everyday interactions with devices such as smartphones and smart speakers (e.g., Amazon Echo smart speaker with Alexa, and Google Home) [13].

Recently, we have begun investigating human-avatar communication within virtual learning environments (virtual simulations and serious games), through the use of intelligent NPCs that incorporate artificial intelligence to allow for more complex (automated) and natural interactive scenarios using speech instead of the traditional dialogue boxes. Our focus is on medical-based virtual learning environments where NPCs (avatars) take on the role of a patient and respond more realistically to any actions taken by the user, who is, in turn, taking on the role of a medical professional (e.g., nurse, or doctor). Although not the focus of our current work, NPCs can also take on the role of a medical professional (e.g., nurse, or doctor) and the user takes on the role of the patient. Our system incorporates eye-tracking technologies and dialogue interpretation to detect and decode user emotions during human (medical professional) –avatar (patient) communication.

Emotions are defined as a subjective response, accompanied by a physiological change that affects behavior and decision making [14]. Using this information, the avatar responds to the user in a realistic manner depending on the detected emotion extracted from the tone of voice. This work is ongoing and, although formal testing is required, preliminary results are promising, indicating the system's ability to detect user emotional states and allow for meaningful communication.

A. Paper Organization

The remainder of the paper is organzied as follows. In Section II, a brief literature review on affective computing (AfC) in virtual environments is provided. Details regarding our system are provided in Section III, while the results a preliminary assessment of our system, final remarks and and plans for future research are provided in Section IV.

II. Background

Affective computing is the interdisciplinary field concerned with the study and development of systems that can recognize, automatically interpret, and process human affects (feelings and emotions) [15]. The goal of affective computing is to design and develop human-user interfaces that respond to the emotional needs of the user naturally, thus bridging the gap between humans and technology [16]. Part of the challenge with this includes ensuring that avatars exhibit adequate human behavior driven by cognitive intelligence, personality, and emotions, all of which create a sense of awareness, and provide an engaging user experience [16]. Believable avatars can facilitate flow [17], enhance haptic interfaces [18], and motivate learning [19]. A fundamental component is the recognition, understanding, and expression of emotions. Motivated by the fact that emotions are involved in many background processes including perception, decision-making, creativity, memory, and social interaction, there is a growing interest in developing a reliable methodology that employs machine learning (ML) to identify the emotional state of a subject [15].

Technologies in graphics, artificial intelligence (AI), sensors, and speech synthesis have advanced, increasing the levels of fidelity and enabling more natural interactions with avatars [20]. Although avatars can be made to appear human, various levels of fidelity can be employed depending on the context of the virtual environment, ranging from cartoon-based characters to photo-realistic ones [21]. Often, the challenge is to ensure that the avatars exhibit adequate human behavior [22]. With respect to medical-based virtual simulations and serious games, avatars (also known as virtual patients or VPs within this context), allow for safe and repeatable exposure to various medical scenarios, providing trainees with the opportunity to practice/learn and make mistakes without real-world consequences [23]. However, given the role of emotion in diagnostics and treatment between both patient and doctor, emotionally intelligent VPs are critical to effective virtual medical-based training [24].

A thorough description of affective computing and intelligent avatars, in general, is beyond the scope of this paper, and an overview of the current state of research in intelligent avatars, with a particular focus on the detection and expression of emotions in medical-based virtual environments an be found in [23]. In addition, an overview of the research surrounding socially intelligent virtual characters and explaining the use of these intelligent virtual characters in serious games is covered in [22]. Finally, a review of psychotherapeutic applications of online avatar technology in [10], outlines how avatars have been used to replace, augment, and facililtate interactions and communication between patient and health professionals while considering the advantages and challenges of implementing this technology within e-mental health interventions [11].

III. System Overview

Our system is comprised of the following sub-systems or modules: i) speech-to-text (STT), tone-analyzer (TA), Assistant, and text-to-speech (TTS). Their role is to manage the communication between the user and the virtual avatar, as presented in Fig. 1. Since our system has been developed specifically for medical-based education, the user will take on the role of the patient, while the avatar will take on the role of the medical professional (e.g., nurse, doctor). However, the system can be applied in other, non-medical applications involving human-avatar communication.

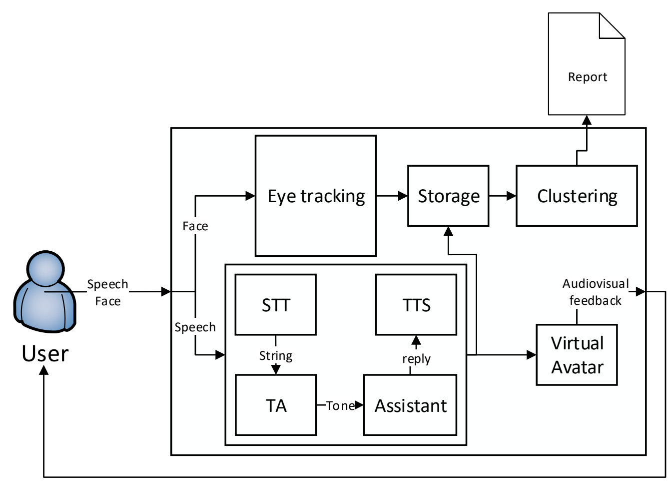

Fig. 1. High-level system architecture.

The interactions with the virtual avatar take place within the virtual environment, which is "navigated" through two main screens. The first screen presents the settings, including the data storage location for the eye-tracking and speech recognition, and enables the configuration of the subject dialogue. The second screen shows the virtual avatar comprised of an interactive computer-generated character, with graphical visual cues providing feedback on the speech-to-text recognition in the form of dialogue triggered by body and speech expressions. The STT feedback shows recognized words so that the user can repeat any word or sentence that was inaccurately detected. Upon completion of the dialogue, a report including four files is created. The files contain information regarding gaze tracking and the dialogue tree produced after the interaction with the avatar. The gaze tracking information is used to generate heat maps during the dialogue, which, in conjunction with the TA, determine the avatar response from the Assistant and the TTS modules.

The Unity game engine was used to implement the system employing its C# scripting API. The user interactions and emotion data collection were achieved employing eye tracking and speech recognition. The eye tracking data was captured employing the Tobii EyeX and the Tobii Unity software development kit (SDK). The Tobii EyeX uses near-infrared light to track the user's eye movements and eye gaze, and operates at 60 Hz. Speech recognition was implemented using IBM Watson services, particularly their speech-to-text for human interactions [25]. Additional layers for interactions included avatar dialogue based on eye tracking using the Scikit-learn library [26], which provides all of the necessary methods to perform unsupervised learning and to produce quantitative measures regarding the user's visual attention. Facial expressions were monitored with the Affectiva software development kit (SDK) [27] that is capable of detecting seven emotion and 20 facial expressions applied to the avatar.

A. *Avatar Development*

The avatar (here, representing the virtual patient), was developed using the Unity Multipurpose Avatar asset in conjunction with the SALSA (Simple Automated Lip Sync Approximation) and the RandomEyes plugins to provide natural lip sync from text-to-speech and eye expressions while talking [28]. Additionally, body animations were added to the avatar to allow for natural body motions during the dialogue, employing the Conversations - Mocap Animation Unity asset. Triggering the animations during the dialogue was achieved by employing IBM Watson's services (i.e., Watson Speech to Text, Watson Text to Speech, Watson Natural Language Understanding, and Watson Tone Analyzer) to ensure that the avatar communication matched the user interactions. IBM Watson runs the STT, TA, Assistant, and TTS modules to let the avatar engage with the dialogue cycle, thus asking and answering questions while triggering non-verbal cues. The IBM Watson plugin employs a release-token system to execute its modules asynchronously, requiring the implementation of C# co-routines instead of requiring updates at every frame which may affect the performance by breaking the fluidity of the dialogue.

B. *Speech-to-Text and Text-to-Speech Cycle*

The primary task of the speech cycle is to transcribe the speech into a string that can be sent to the TA module. The TA extracts the tone, interruptions, and silence gaps in the dialogue, which are sent to the Assistant for generating the response to the user. When active, the STT module records the user's voice including any silence between words. After recognizing the words, the dialogue is fragmented depending on how the silence threshold was configured, thus changing punctuation for example when being transcribed.

Using these assembled strings (including punctuation), the TA module determines the tone and emotion associated with the user verbal communication when speaking to the avatar. Then the TA module creates a file containing a list of tones with values classifying them into anger, fear, joy, sadness, analytical, confident, and tentative based on IBM Watson processing. The detected tone value is then used to trigger facial animations on the virtual avatar and thus allow for enhanced empathy during the dialogue. The vocal tones were classified into two emotion categories as positive and negative to determine the facial expression response from the avatar. The positive emotion comprises tones of joy and confidence, while the negative is based on sadness, fear, and anger. Since the plugin does not support neutral tones, a JavaScript Object Notation (JSON) function was implemented to set a neutral expression on the avatar when the positive and negative scores are equal.

C. *The Assistant*

The IBM Watson Assistant module allows for the creation of more realistic and natural dialogue interaction through the user's voice (where here, the user is taking on the role of a medical professional), to communicate with the avatar. The Assistant module relies on machine learning to learn how to respond to the various voice inputs from the user. The training of the ML component relies on two elements, the intents and the entities. The intents represent the requests made by the user while interacting with the Assistant. The requests are created based on the dialogue and, for each request, an intent is created. For example, the string "hours bank" can create an association with a bank operating hours. The contextual entities component allows identifying interesting parts of the dialogue such as names, dates and places. In conjunction with Watson system entities, the Assistant is able to define the entities with synonyms and fuzzy matching. Fuzzy matching provides the opportunity to overcome potential grammatical errors from the STT module. The intent and entities create a dialogue tree, where slots are filled-in based on the user dialogue input, and correspondence between the terms is found through the fuzzy matching feature of the plugin. Finally, the TTS module employs the speech synthesis markup language, providing a standard way to generate synthetic speech that includes expressiveness, voice transformation, and customization for pronunciation.

D. *Eye Tracking*

Gathering information (metrics) related to areas of focus during the virtual avatar dialogue was achieved in two stages: i) while the user is speaking, and ii) while the user is listening to the avatar speak. The gaze information was used to create a visual representation of the collected data in the form of a heat map that allows visualizing the areas of focus during the dialogue. Fig. 2 illustrates an example scene with the virtual avatar, the heat map overlaid on it (e.g., red regions indicate areas where the focus was higher), and dialogue strings.

E. *Clustering*

The Density-based spatial clustering of applications with noise (DBSCAN) [29], k-means [30], and mean shift [31] clustering algorithms were implemented to compute quantitative measures from the recorded eye movements. The goal is to group the eye data coordinates on the screen with similar areas of fixation, as their proximity can provide insights on why the gaze was attracted to a specific location. Fig. 2 illustrates several clusters that can be associated with the semantic context of the dialogue to determine their relevance during the interactions.

Once the areas of interest have been defined, fixation or derived metrics can be used to obtain information about the number of fixations, fixations per area, duration, spatial density, and time to fixate to the target. Saccade-derived metrics can also be extracted, based on the quick, simultaneous movement of both eyes between two or more phases of fixation in the same direction. Additionally, the ratio between the number of fixations and saccades indicates the amount of time taken to navigate on-screen information (e.g., a higher ratio indicates a higher searching time).

Fig. 2. Virtual avatar during the interactions with overlaid heat map clusters (red regions). The heat map represents where the user focused the most when talking to the avatar..

IV. Final Remarks

The current availability of affordable computational processing tools along with the accompanying decrease in the size of electronic components has led to the availability of consumer-level immersive technologies and considerable advancements in affective computing and artificial intelligence. The technology growth associated with the development of intelligent human-avatar communications has led to innovative human-computer interactions employing body tracking, physiological responses, and data processing using AI.

In our effort to develop more realistic virtual simulations and serious games for medical education, health and well-being, here we presented a prototype capable of detecting a user's emotions and thus allow for realistic and meaningful human-avatar communication. Through the use of eye-tracking and speech interpretation, we seek to depart from traditional on-screen pre-defined dialogue trees that cannot capture emotion user emotions. Although our system will be applied to medical-based virtual learning tools (e.g., virtual simulations, and serious games) where the avatar takes on the role of a patient and the user takes on the role of a medical professional (e.g., nurse, doctor), our system is applicable to any application involving human-avatar communication.

A preliminary assessment of the tool was designed to sustain a short dialogue with a virtual avatar leading the conversation. The dialogue requires the user to answer the questions starting with name and occupation, followed by questions about color preference, current activities, and enjoyable activities. To assess how the system responds, five participants with computer development background volunteered to interact with the virtual avatar. Before starting the dialogue, both the eye tracker and microphone were calibrated to guarantee adequate tracking and interactions. The calibration accounts for unique participant traits, including height, interpupillary distance, and distance from the screen, although these traits may affect eye tracker performance. In terms of speech recognition, the calibration allows the system to account for the participants' pitch and accent. Two English native speakers and three non-English native speakers whose first language was Italian and Spanish interacted with the virtual avatar. Although all users completed the dialogue, there were significant delays caused by the IBM Watson plugin while sending and receiving the STT data. Since the system did not have any previous training, all users had to repeat themselves several times with some words, thus causing a sense of frustration for all. The avatar responses based on the emotion detected from the speech triggered the avatar animations effectively. However, the animation, in conjunction with TTS, resulted in an uncanny valley dialogue.

As mentioned, the system presented here is preliminary and significant work remains for the future. Notwithstanding, the current implementation demonstrated the potential and promise of using ML as a tool to develop more natural interactions. At this point, we acknowledge challenges associated with appropriate avatar animation and the need to explore other cloud-based STT and TTS solutions to reduce latency and improve speech recognition accuracy. As expected, the system had more difficulties understanding words from non-English speakers, increasing the dialogue time and making it harder to complete. Formal testing of the method will be conducted to quantify its accuracy in the near future. The method may be adjusted/updated depending on the experimental outcomes and will then be incorporated in two virtual simulations developed for intensive short-term dynamic psychotherapy (ISDP), a form of psycho-dynamic psychotherapy that provokes emotional experiences from the patient to facilitate a corrective emotional experience (see [32]).

Acknowledgment

The authors acknowledge the funding to support this work from ERASMUS+ in the form of a scholarship to S. Cinieri, the *Natural Sciences and Engineering Research Council of Canada* (NSERC), and the *Social Sciences and Humanities Research Council of Canada* (SSHRC). Work was carried out in the context of the VR@POLITO initiative of Politecnico di Torino.

References

[1] D. Wortly. The future of serious games and immersive technologies and their impact on society. In Y. Baek, R. Ko, and T. Marsh (Eds.), *Trends and Applications of Serious Gaming and Social Media*, Springer Science+Business Media Singapore, 2014.

[2] C. Anthes,, R. J. García-Hernández, M. Wiedemann, and D. Kranzlmüller. State of the art of virtual reality technology. In *Proceedings of the 2016 IEEE Aerospace Conference*, pp. 1-19, 2016.

[3] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba. Eye tracking for everyone. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2176-2184, 2016.

[4] J. Titz, A. Scholz, and P. Sedlmeier, P. Comparing eye trackers by correlating their eye-metric data. *Behavior Research Methods* 50(5), 1853-1863, 2018.

[5] S. Hall, and Takabashi. Augmented and virtual reality: the promise and peril of immersive technologies. World Economic Forum, https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/augmented-and-virtual-reality-the-promise-and-peril-of-immersive-technologies/ (2017) last accessed 3 Jan 2020.

[6] G. Huff, and B. Saxberg. Full immersion—How will 10-year-olds learn? *Education Next* 9(1), 79–82, 2009.

[7] R. Shewaga, A. Uribe-Quevedo, B. Kapralos, and F. Alam. A comparison of seated and room-scale virtual reality in a serious game for epidural preparation. *IEEE Transactions on Emerging Topics in Computing, Special issue on Innovation in Technologies for Educational Computing*, DOI: 10.1109/TETC.2017.2746085, 2018.

[8] S. Valdivia, R. Blanco, A. Uribe-Quevedo, L. Penuela, D. Rojas, and B. Kapralos. Development and evaluation of two posture tracking user interfaces for occupational health care. *Advances in Mechanical Engineering* 10(6), 1-12, 2018.

[9] A. Uribe-Quevedo, and B. Kapralos. Shoulder rehabilitation exergaming. In A. Brooks, S. Branham, B. Kapralos, and L. Jain (Eds.) *Recent Advances in Technologies of Inclusive Well-Being: Wearables, Virtual Interactive Spaces (VIS)/Virtual Reality, Emotional Robots, Authoring tools, and Games (Serious/Gamification),* Springer Intelligent Systems Reference Library Series, Ch. 6, pp. 105-125, 2017.

[10] I. C. Rehm, E. Foenander, K. Wallace, J. M. Abbott, M. Kyrios, and N. Thomas. What role can avatars play in e-mental health interventions? Exploring new models of client–therapist interaction. *Frontiers of Psychology*, November, 2016 | https://doi.org/10.3389/fpsyt.2016.00186.

[11] A. Gorini, A. Gaggioli, C. Vigna, and G. A. Riva. A second life for eHealth: prospects for the use of 3-D virtual worlds in clinical psychology. *Journal of Medical Internet Research*, 10(3):e21, 2008.

[12] D. Shin. Empathy and embodied experience in virtual environment: To what extent can virtual reality stimulate empathy and embodied experience? *Computers in Human Behavior* 78(1): 64-73, 2018.

[13] F. Habler, V. Schwind, & N. Henze. Effects of Smart Virtual Assistants' Gender and Language. In Proceedings of the *ACM Mensch und Computer Computer*, pp. 469-473, 2019.

[14] L. A. Lefton. Psychology. 5th ed. Boston: Allyn and Bacon, 1994.

[15] J. R. Marín-Morales, J. L. Higuera-Trujillo, A. Greco, J. Guixeres, C. Llinares, E. P. Scilingo, M. Alcañiz and G. Valenza. Affective computing in virtual reality: Emotion recognition from brain and heartbeat dynamics using wearable sensors. *Scientific Reports*, 8:13657, 2018. DOI:10.1038/s41598-018-32063-4.

[16] P. De Byl. A conceptual affective design framework for the use of emotions in computer game design. *Cyberpsychology* 9(3), Article 4, 2015.

[17] D. Johnson, and J. Wiles. Effective affective user interface design in games. Ergonomics, 46(13-14), 1332-1345, 2006.

[18] E. Hudlicka. Affective computing for game design. In *Proceedings of the 4th International North American Conference on Intelligent Games and Simulation*, pp. 5-12, 2008.

[19] J. Robison, S. McQuiggan, and J. Lester. Evaluating the consequences of affective feedback in intelligent tutoring systems. In *Proceedings 3rd International Conference on Affective Computing and Intelligent Interaction*, pp. 1-6, 2009.

[20] S. Lee, G. Carlson, S. Jones, A. Johnson, J. Leigh, and L. Renambot. Designing an expressive avatar of a real person. In: J. Allbeck, N. Badler, T. Bickmore, C. Pelachaud, and A. Safonova (Eds.) Intelligent Virtual Agents. IVA 2010. Lecture Notes in Computer Science, vol 6356. Springer, Berlin, Heidelberg, pp. 64-76, 2010.

[21] Y. Jiang, E. E. O'Neal, J. P. Yon, L. Franzen, P. Rahimian, J. M. Plumert, and J. K. Kearney. Acting together: Joint pedestrian road crossing in an immersive virtual environment. *ACM Transactions on Applied Perception* 15(2), 8, 8:1--8:13, 2018.

[22] N. Magnenat-Thalmann, and Z. Kasap. Virtual humans in serious games. *Proceedings of the International Conference on CyberWorlds*, pp. 71-79, 2009.

[23] B. Kapralos, A. Uribe-Quevedo, K. Collins, and A. Dubrowski. Intelligent avatars and emotion in medical-based virtual environments. *Intelligent Decision Technologies*, 13(4), 2019.

[24] D. Kozlowski, M. Hutchinson, J. Hurley, J. Rowley, and J. Sutherland. The role of emotion in clinical decision making: an integrative literature review. *BMC Medical Education* 17(1), 2017. doi:10.1186/s12909-017-1089-7

[25] S. Varrasi, A. Lucas, A. Soranzo, J. McNamara, and A. Di Nuovo. IBM cloud services enhance automatic cognitive assessment via human-robot interaction. In: *New Trends in Medical and Service Robotics*, pp. 169-176, Springer, 2019.

[26] O. Kramer. Scikit-learn. In Machine learning for evolution strategies, pp. 45-53, Springer, 2016.

[27] M. Magdin, and F. Prikler. Real time facial expression recognition using webcam and SDK affectiva. *International Journal of Interactive Multimedia and Artificial Intelligence*, 5(1), 7-15, 2018.

[28] P. Zikas, V. Bachlitzanakis, M. Papaefthymiou, S. Kateros, S. Georgiou, N. Lydatakis, and G. Papagiannakis. Mixed reality serious games and gamification for smart education. *Proceedings European Conference on Games Based Learning*, pp. 805, 2016.

[29] K. Naqshbandi, T. Gedeon, and U. A. Abdulla. Automatic clustering of eye gaze data for machine learning. *Proceedings 2016 IEEE International Conference on Systems, Man, and Cybernetics* (SMC), pp. 1239-1244. 2016.

[30] C. R. Latimer. Eye-movement data: Cumulative fixation time and cluster analysis. *Behavior Research Methods, Instruments, & Computers* 20(5):437-470, 1988.

[31] A. Santella, and D. DeCarlo. Robust clustering of eye movement recordings for quantification of visual interest. In *Proceedings of the 2004 Symposium on Eye Tracking Research & Applications*, pp. 27-34, 2004.

[32] K. Wilcocks, B. Kapralos, A. Kalpin, D. Chandross, and A. Abbass. A dialogue intervention simulation framework to facilitate psychotherapy training. In *Proceedings IEEE International Symposium on Consumer Technologies*, pp. 1-2, 2019.