A GPT-based Practical Architecture for Conversational Human Digital Twins

Shahan Mehtab Iqbal[1,*], Matias Volonte[1], Bart Knijnenburg[1] and Nina Hubig[1]

[1]*Clemson University, USA*

Abstract

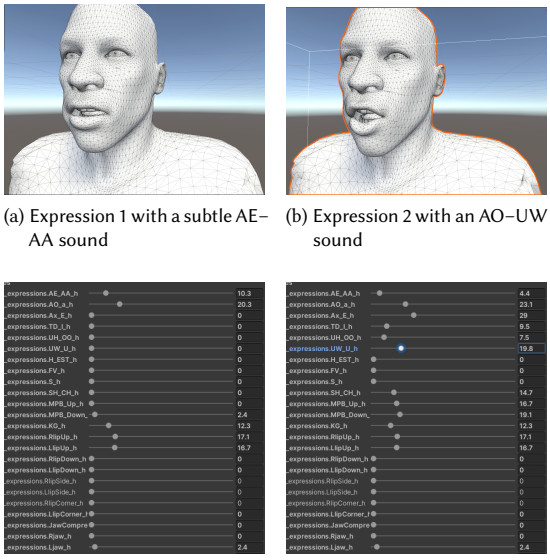
Virtual human twins, also known as digital clones, are computer-generated replicas of human beings that can be used for a variety of purposes, such as entertainment, education, and healthcare. The vision of creating virtual human twins with distinct personalities and emotional responses has the potential to revolutionize these industries by enabling more personalized and engaging experiences for users. This research aims to explore this vision under the aspect of responsponsible artificial intelligence (AI) and identify the tools, technologies, and limitations involved in its realization. We will examine the current state of virtual human twin technology and evaluate the open-source resources available for building them. Additionally, we will investigate the requirements of virtual human twins to be created in a responsible way and analyze the challenges that need to be overcome to create more sophisticated versions. The findings of this research will provide valuable insights into the future of virtual human twin technology and inform future efforts in this field.

Keywords

digital clone, virtual human twin, chatgpt, architecture avatar, trustworthy ai

1. Use Case

Human digital twins (HDTs) gained popularity in recent years and play an essential role in the expressiveness of oneself in the virtual world[1]. Often, an HDT is defined as the virtual clone or replica of the user, that in the context of machine learning can behave autonomously and has all human channels (lingual, emotional etc.) to connect with humans in a mimicking way [2]. The use case now is defined as creating a human digital twin as a minimum viable product with open source possibilities including concepts of responsible and ethical artificial intelligence (AI). While technology has to cover many more grounds before full-fledged HDTs can be used in the real world, the current state of artificial intelligence has demonstrated convincing performance within contextualized environments. With the advent of GPT-3s extension, ChatGPT and similar architectures [3] can hold a plausible written conversation with a human, even memorizing the previous questions asked. Speech-to-text and text-to-speech natural language processing allows for accurate, yet semi-emotional transfer of voice and intonation. Additionally, advanced game engines such as MetaHumans [4], provide to HDTs the capabilities to produce believable representations of humans. Including conversational behaviors such as facial animation, lip sync, and emotions.



(a) Expression 1 with a subtle AE–AA sound

(b) Expression 2 with an AO–UW sound

(c) Paramters for expression 1 above

(d) Parameters for expression 2 above

Figure 1: Granular blend shapes controlled programmatically by speech synthesis parameters synchronized with audio timesteps to generat lip movement animation during speaking.

AiOfAi'23: 3rd Workshop on Adverse Impacts and Collateral Effects of Artificial Intelligence Technologies, Macao, China

*Corresponding author.

mehtabi@clemson.edu (S. M. Iqbal); mvolont@clemson.edu (M. Volonte); bartk@clemson.edu (B. Knijnenburg); nhubig@clemson.edu (N. Hubig)

0000-0002-0877-7063 (S. M. Iqbal); 0000-0001-7116-9338 (M. Volonte); 0000-0002-9421-8566 (B. Knijnenburg); 0000-0002-9421-8566 (N. Hubig)

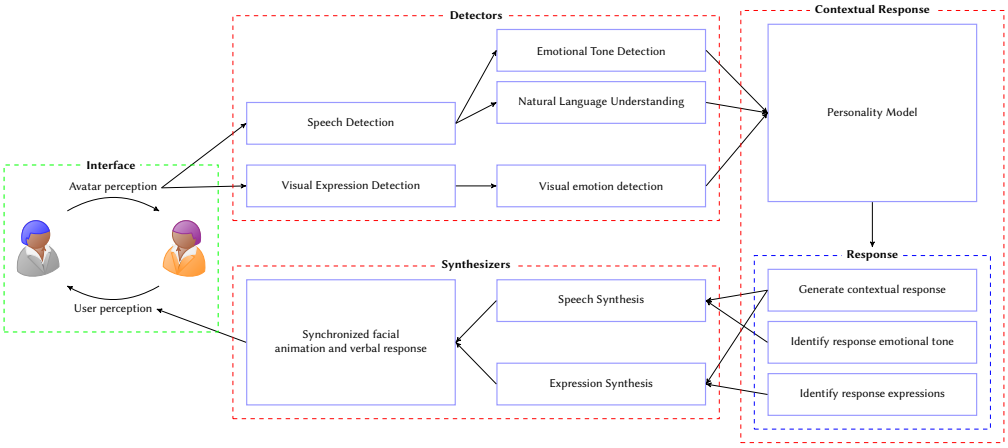


Figure 2: The flow diagram of the top-level components that make up the HDT virtual conversational agent.

2. Architecture

We will focus on the software elements of the system and leave it in good faith the obviousness of requiring necessary hardware, such as a microphone and camera as input tools and a display unit and speakers as output devices. With that out of the way, the system can be broadly divided into three categories 1) detectors, 2) contextual response generator, and 3) synthesizers. The detectors extract speech content from the microphone feed and facial expressions or other nonverbal cues from the camera feed. The contextual response generator consists of language models, computer vision algorithms, and audio processors to contextualize a response to the HDT personality. Finally, the synthesizers generate the relevant verbal response and facial animation that is reflected on the avatar. Figure 2 shows this architecture with its three subsystems and the interface.

At a minimum, the HDT presented thus far can be implemented in phases as long as they are the right corresponding components from each of the three subsystems. A minimal conversational agent can be implemented with only the speech detector, contextual speech response model, and speech synthesizer. Ideally though, each aspect is fully integrated with more optional components like expression detections to achieve more realistic results.

3. Implementation

The implementation of a minimal HDA conversational agent uses a game engine such as Unity3D [5], DeepSpeech[6], and a transfer-learned GPT-3[3]. Additionally, we will also demonstrate how each component can be incrementally improved or integrated to realize a full-fledged HDT conversational agent. Unity3D is used as an interface for user interaction that also acts as the client that sends audio clips to a server and consumes the server response of response audio clip and synchronized mouth animation (figure 1).

LLMs have made significant progress in the field of natural language processing. GPT-3 is a language model that can be used to generate text responses to a given input. The model is trained on a large corpus of text. So, we use a corpus of conversation in some pre-defined scenario to train a language model as the contextual speech response model. The transfer learning provides the basis for our personality model, that is, the personality we want the avatar to display.

Using a text based language model necessitates the speech input received by the conversational agent to be converted to text. This is done using a speech-to-text engine such as DeepSpeech [6]. DeepSpeech is an open source speech-to-text engine that uses a deep neural network to convert speech to text. The model is trained on a large corpus of speech and text. However, real-time streaming speech-to-text is not supported by DeepSpeech. So, we use an amplitude thresholding algorithm to detect the start and end of a speech utterance. The speech utterance segment is then sent as a clip to the server for processing. The server uses the DeepSpeech engine to convert the speech to text and then uses the language model to generate a response.

4. Digital Twins and Ethics

Digital twin technologies can be harnessed for both positive and negative purposes. Ethical considerations encompass evaluating potential dual-use scenarios and establishing safeguards against misuse.

Digital twins thrive on vast datasets gathered from sensors and sources embedded within the physical entities they emulate [7, 8]. This flow of data, while enabling accurate modeling and decision-making, raises significant privacy concerns. The data collected often comprises personal, sensitive, and proprietary information, necessitating stringent measures to prevent unauthorized access and potential misuse[9].

Respecting user privacy involves obtaining informed consent for data collection, storage, and utilization in the digital twin ecosystem. Users should be well-informed about the purposes and scope of data usage, and their consent should be obtained in a transparent and comprehensible manner. Empowering users to control the extent to which their data is shared and used is essential to maintaining trust and upholding ethical standards [10].

Automated decision-making through digital twins introduces a set of ethical challenges[11]. The potential for biases, unexpected outcomes, or systemic failures mandates a proactive approach to identify and address these issues. Ensuring that accountability extends to the identification and mitigation of unintended consequences is vital for responsible deployment[12].

References

[1] R. Hooi, H. Cho, Being immersed: avatar similarity and self-awareness, in: Proceedings of the 24th Australian Computer-Human Interaction Conference, 2012, pp. 232–240.

[2] H. Lonsdale, G. M. Gray, L. M. Ahumada, H. M. Yates, A. Varughese, M. A. Rehman, The perioperative human digital twin, Anesthesia & Analgesia 134 (2022) 885–892.

[3] L. Floridi, M. Chiriatti, GPT-3: Its Nature, Scope, Limits, and Consequences, Minds and Machines 30 (2020) 681–694. doi:10.1007/s11023-020-09548-1.

[4] E. Games, High-fidelity digital humans made easy, 2023. URL: https://www.unrealengine.com/en-US/metahuman, last accessed 16 February 2023.

[5] J. Xie, Research on key technologies base unity3d game engine, in: 2012 7th international conference on computer science & education (ICCSE), IEEE, 2012, pp. 695–699.

[6] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, et al., Deep speech: Scaling up end-to-end speech recognition, arXiv preprint arXiv:1412.5567 (2014).

[7] A. El Saddik, Digital twins: The convergence of multimedia technologies, IEEE multimedia 25 (2018) 87–92.

[8] A. El Saddik, F. Laamarti, M. Alja'Afreh, The potential of digital twins, IEEE Instrumentation & Measurement Magazine 24 (2021) 36–41.

[9] J. Lane, C. Schur, Balancing access to health data and privacy: a review of the issues and approaches for the future, Health services research 45 (2010) 1456–1467.

[10] S. B. Far, A. I. Rad, Applying digital twins in metaverse: User interface, security and privacy challenges, Journal of Metaverse 2 (2022) 8–15.

[11] M. B. Van Riemsdijk, C. M. Jonker, V. Lesser, Creating socially adaptive electronic partners: Interaction, reasoning and ethical challenges, in: Proceedings of the 2015 international conference on autonomous agents and multiagent systems, Citeseer, 2015, pp. 1201–1206.

[12] B. Mittelstadt, Near-term ethical challenges of digital twins, Journal of medical ethics 47 (2021) 405–406.