Kashihara, A. et al. (Eds.) (2024). Proceedings of the 32nd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Technology Considerations in Building Virtual Educational Avatars

Antun DROBNJAK & Ivica BOTICKI*
Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia
*ivica.boticki@fer.hr

Abstract: This paper examines the area of technology development for virtual educational avatars. Since such avatars can be real-like and closely resemble real or fictional persons, the paper examines practical aspects of such designs, linking technology designs to pedagogy requirements. Following these, technology development considerations for realistic voice and video generation and the corresponding architecture which enables realistic virtual avatar for use in education are discussed. Additionally, using powerful CPUs and GPUs shows how current technology allows for efficient and sophisticated video and audio generation, emphasizing the need for optimized resource management to balance processing speed and output quality.

Keywords: avatars, personas, virtual, education, technology design

1. Introduction

Generative artificial intelligence brings a new dimension to the field of technology use in education, as it enables the creation of virtual characters that imitate real people, historical figures, and even fictional characters. Such a design potentially brings a new pedagogically innovative and interactive learning dimension, and in combination with visual models and sound synthesis can contribute to a realistic representation of an educational virtual character.

The emergence of generative artificial intelligence re-intensifies the discussion on the use of artificial intelligence to improve education (Baidoo-Anu & Owusu Ansah, 2023). Generative artificial intelligence is beginning to be an integral element of interactive tools and systems. Such interactive systems have the ability to improve the way students engage with educational content and access learning resources. The adaptability of generative AI models has great potential in supporting personalized learning and improving interaction in educational processes, which could lead to better education and learning (Xu et al., 2022).

To create realistic educational virtual characters, it is essential to develop concepts for both image modeling and sound synthesis. Image models will shape the characters' appearances, while sound synthesis will generate lifelike voice experiences, adapting to linguistic properties. These characters can then guide students through educational processes, using AI to tailor interactions to individual needs. However, an additional challenge is supporting multiple spoken languages, which requires research into sound synthesis that accommodates specific phonetics, intonations, and speech rhythms.

This paper begins with a theoretical background, laying the foundation for understanding the subsequent developments in persona interaction and virtual avatar creation. Following this, it presents a general model overview. The article then delves into the details of video and audio generation, explaining the processes and technologies involved. Building on this, it introduces an integration of these elements to create a seamless virtual avatar experience. Finally, the article showcases the performance results, demonstrating the efficiency of the proposed system.

2. Theoretical Background

Haller & Rebedea (2013) present a method for building a conversational agent that has personality and knowledge about historical figures and can be used in an educational context. Likewise, Fu et al. (2022) point out that the memory property of certain agents in a conversation can significantly improve the answers they give. Xu et al. (2022) propose an improved framework for designing pedagogical agents, which can ask questions, give feedback, are based on the scaffolding principle and dynamically adapt to the students.

Artificial intelligence algorithms can analyze vast amounts of data from social media, online behavior, and surveys to identify patterns and trends that define the characteristics of individuals as well as notable historical figures. This information can then be integrated into mathematical models to generate 2D or 3D avatars that visually represent the "essence" of these virtual characters. These avatars can engage in realistic and contextually relevant interactions with students, offering them guidance, explanations, and answers to questions. The results of related studies (O. L. Liu, 2012; Spinath, 2009) suggest that the way a student relates to a teacher can have a significant impact on a student's attitude and level of motivation. Therefore, it is important to investigate and consider how AI-generated teachers could be used to increase motivation in online learning (Kosmyna, 2020).

The concept of creating image models for visual representations of virtual characters has enormous potential to improve the educational experience by adapting content to the individual needs and preferences of students (Pataranutaporn, 2021; Chiu, 2021). Given that modern learning platforms use artificial intelligence algorithms to analyze student performance and behavior, and enable the delivery of customized content (Wei, 2021), the integration of image models into these platforms can create realistic virtual characters that guide students through educational processes and simultaneously analyze their tendencies and progress.

The introduction of generative adversarial networks (GANs) (Goodfellow, 2014) has enabled the realistic synthesis of digital content, including the generation of photorealistic images, voice cloning, facial animation, and image translation from one form to another (Mirsky, 2021; Karras, 2020).

In the modern age of digital technology, voice synthesis is becoming a key component of many applications, including virtual assistants, interactive learning, and entertainment. Synthesizing voice is not just a matter of reproducing sound waves, but an artistic and technological achievement that requires a combination of advanced algorithms and precise modeling. In this context, three models: Wavenet, Tacotron2 and VITS - become central in the development of sound models for the sound representation of virtual characters.

Together, these technologies enable the development of innovative solutions for voice synthesis, which provide precision and quality and enable the personalization and flexibility. In the context of education, technology allows students to experience a lesson through the voice of a famous person, which is often more engaging than learning from a textbook.

3. A Model for Realistic Virtual Avatars

The creation of an immersive virtual educational persona is a multi-faceted process that integrates various advanced technologies and theoretical frameworks to deliver an engaging and personalized learning experience. This section provides an overview of the model flow, starting from the inputs to the final persona that communicates with students, see Figure 1.

The process begins with the integration of open-source and large language models. These models are the foundation for generating synthesized, realistic voices that sound natural and enhance the authenticity of the interaction. Simultaneously, precise image models are employed to create visual representations of the persona. These models are designed to respond dynamically to student interactions. By leveraging advanced image processing and animation techniques, the persona can exhibit facial expressions and gestures that correspond to the spoken content, further enhancing interaction realism.

The development of the virtual educational persona is guided by a conceptual pedagogical framework that incorporates insights from learning sciences and pedagogy. This

theoretical approach ensures that the educational strategies employed by the persona are grounded in proven educational principles, enhancing learning effectiveness.

Once the inputs are processed through this framework, the result is an immersive virtual educational persona. This persona is tailored using both pedagogical and personalization techniques, making it capable of adapting to the individual needs and preferences of each student. The personalized approach ensures that the content delivered is relevant and suitable for the learner's current state and learning objectives.

Finally, the virtual educational persona interacts with students in a way that mimics human interaction, fostering an engaging and supportive learning environment. In summary, the general model overview highlights a comprehensive and integrated approach to creating an immersive virtual educational persona. By combining advanced language models, precise image models, IoT technology, and robust pedagogical frameworks, the model delivers a highly engaging and personalized learning experience that interacts dynamically with students to enhance their educational journey.

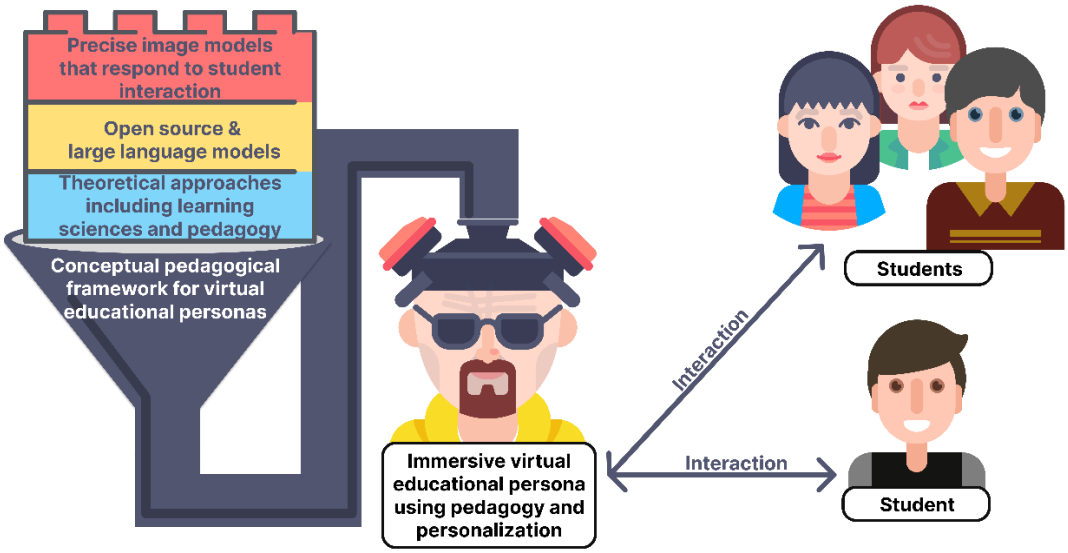

Figure 1 Model overview

3.1 Video Generation

Video generation technology has advanced significantly, enabling the creation of lifelike avatars from still images. One such groundbreaking method is MakeItTalk (Yang Zhou et al., 2020), which brings static portraits to life by transforming them into expressive talking heads. Using a single portrait and an audio clip, MakeItTalk generates realistic animations that synchronize facial expressions and head movements with the spoken audio. This deep learning-based approach effectively disentangles speech content from speaker identity, ensuring that the lip movements and facial dynamics are both accurate and personalized to the individual depicted in the portrait.

MakeItTalk leverages advanced neural networks to achieve high fidelity in speech and facial synchronization. The system maps the audio features to corresponding facial movements, creating a seamless and natural-looking talking head. This technology can animate various facial expressions, from subtle eyebrow raises to complex lip movements, providing a rich and engaging user experience (Zhou et al., 2020). The result is a compelling and interactive avatar that can be utilized across multiple applications.

3.2 Sound Generation

Text-to-speech (TTS) synthesis is the technology that converts written text into spoken words. It has a wide range of applications, from virtual assistants and audiobooks to accessibility tools for the visually impaired (Dheeraj Jalali, 2020). Traditional TTS systems often rely on concatenative synthesis or parametric synthesis (Dutoit, 1997). However, modern TTS has

significantly advanced with the advent of deep learning, leading to models like Tacotron 2 and WaveNet, which generate natural-sounding speech (Kim et al., 2020).

Glow-TTS stands out for its simplicity in training, speed in synthesis, and robustness in performance. It opens up new possibilities for real-time TTS applications and can be further developed for tasks like prosody transfer or style modeling. The model's source code and audio samples are publicly available (Coqui GmbH, 2020), encouraging further research and development in the field of speech synthesis. By pushing the boundaries of TTS technology, Glow-TTS contributes significantly to the evolution of more natural speech synthesis.

3.3 Virtual Avatar with Video and Sound Synchronization

Advancements in video and sound generation technologies have paved the way for the creation of virtual avatars with unprecedented realism and interactivity. By seamlessly integrating video and sound synchronization techniques, it is now possible to bring digital characters to life, bridging the gap between static images and dynamic speech.

The convergence of video generation, such as MakeItTalk (Yang Zhou et al., 2020), and sound generation, like Glow-TTS (Coqui GmbH, 2020), presents a remarkable opportunity to create virtual avatars that not only look realistic but also sound authentic. By harnessing the power of deep learning algorithms, these technologies enable the synthesis of lifelike facial expressions and natural-sounding speech, respectively.

The process of creating a virtual avatar with synchronized video and sound involves several steps. First step begins with input data, which typically consists of a still image or portrait of the avatar, along with the corresponding text that the avatar will speak. The text can either be predefined, provided by the user, or generated using another model capable of text generation. The second step is to convert written text into spoken words using advanced text-to-speech synthesis models like Glow-TTS. While, in third step still image or portrait of the avatar is fed into a video generation model such as MakeItTalk. This model animates the image, creating a realistic talking head that synchronizes facial movements with the audio.

Once the integrated video and sound generation process is complete, the output is a video file containing synchronized audio sound. This video file features a virtual avatar, animated based on the provided still image or portrait, speaking the generated or predefined text with corresponding lip movements and facial expressions.

4. Evaluating Model for Virtual Avatars

The demand for high-quality, automated avatar generation has grown significantly in recent years. This process involves computationally intensive audio and video generation. To understand the efficiency of different computational setups, performance tests were conducted. Specifically, the execution times of generating avatars with varying sentence lengths were evaluated on a CPU, a single GPU, and dual GPUs, see Table 1. The hardware configurations used for the tests are: CPU AMD Ryzen 5 5625U with Radeon Graphics, running at 2.30 GHz and GPU NVIDIA A100-SXM4-40GB.

4.1 Results

The tests measured the time taken to generate avatars with audio and video for 1, 4, and 8 sentences. The execution times are summarized below.

Table 1. *Performance test*

Number of Sentences	Length of Audio/Video	CPU Time	1 GPU Time	2 GPUs Time
1 sentence	16s	510s	60s	26s
4 sentences	48s	1013s	131s	54s
8 sentences	96s	N/A	503s	105s

4.2 Analysis

For a single sentence of 16 seconds in length, the CPU took 510 seconds to complete the task, indicating that CPU-based processing for avatar generation is feasible but not optimal for real-time applications. When a single GPU was used, the time reduced significantly to 60 seconds, demonstrating the processing power of the GPU. With two GPUs, the time further decreased to 26 seconds, showing a substantial improvement in processing efficiency.

For four sentences of 48 seconds length, the CPU time increased to 1013 seconds. For a single GPU, the time was 131 seconds, which is a reasonable increase from the single sentence generation, maintaining the efficiency advantage over the CPU. The dual GPU setup completed the task in 54 seconds, continuing to demonstrate significant time savings.

For eight sentences of 96 seconds in length, the CPU was unable to complete the task within a practical timeframe. The single GPU took 503 seconds, showing an increase that reflects the complexity and size of the task. The dual GPU configuration handled the task in 105 seconds, maintaining its trend of superior performance and scalability.

The performance tests clearly illustrate the advantages of using GPUs over CPUs for avatar generation involving TTS and video synthesis. While the CPU struggled with longer tasks, GPUs proved to be significantly more efficient. This scalability and reduction in processing time are crucial for applications requiring near-real-time avatar generation.

5. Discussion

The integration of video and sound generation technologies has revolutionized the creation of virtual avatars, offering seamless synchronization of audio with animated visuals. While this advancement presents challenges related to resource requirements, it also opens up opportunities to enhance user experience.

Generating synchronized video with audio sound demands significant computational resources, including high-performance GPUs, ample memory, and processing power. Additionally, scaling up operations to accommodate increased workload or user base may require careful consideration of hardware scalability and optimization strategies.

Despite the resource challenges, there are opportunities for optimization to enhance efficiency and reduce computational overhead. Advances in hardware technology, such as improvements in GPU architecture and memory optimization techniques, contribute to more efficient video and sound generation processes. Algorithmic optimizations and parallel processing techniques further improve performance and scalability, enabling more cost-effective and scalable solutions.

Achieving a balance between performance and cost is essential when deploying synchronized video with audio sound solutions. Organizations must carefully evaluate their resource requirements and budget constraints to determine the most cost-effective approach. Optimization strategies, such as hardware utilization optimization and efficient algorithm implementation, help strike the right balance between performance and cost.

By leveraging synchronized video with audio sound technologies, organizations can create more human-like and relatable virtual avatars, leading to higher user satisfaction and improved interaction quality. Ultimately, enhancing user experience remains a primary goal, driving innovation and advancement in synchronized video and sound generation technologies.

6. Conclusions

The integration of video and sound generation technologies has revolutionized virtual avatar creation by enabling seamless synchronization of audio with animated visuals. While significant computational resources such as high-performance GPUs and ample memory are required, the opportunities for optimizing performance and enhancing user experience outweigh these challenges. Investments in robust hardware infrastructure and optimization strategies can make these technologies more efficient and cost-effective. Advances in GPU architecture, memory optimization, and algorithmic improvements can help address scalability

and performance issues. The most significant benefit of this technology is the enhanced user experience it offers. Lifelike virtual avatars capable of seamless audio-visual synchronization foster deeper engagement and satisfaction across various applications, including virtual assistance, digital storytelling, education, and customer service. By prioritizing user experience and continuing to refine these technologies, organizations can harness their full potential, creating compelling and immersive digital interactions that bridge the gap between digital and real-world communication.

Acknowledgements

We would like to thank all the people who prepared and revised previous versions of this document.

References

Baidoo-Anu, D., & Owusu Ansah, L. (2023). Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4337484

Chiu, T. (2021). Student engagement in k-12 online learning amid covid-19: A qualitative approach from a self-determination theory perspective. Interactive Learning Environments, 1-14.

Fu, T., Zhao, X., Tao, C., Wen, J. R., & Yan, R. (2022). There Are a Thousand Hamlets in a Thousand People's Eyes: Enhancing Knowledge-grounded Dialogue with Personal Memory. Proceedings of the Annual Meeting of the Association for Computational Linguistics, 1. https://doi.org/10.18653/v1/2022.acl-long.270

Haller, E., & Rebedea, T. (2013). Designing a chat-bot that simulates an historical figure. Proceedings - 19th International Conference on Control Systems and Computer Science, CSCS 2013. https://doi.org/10.1109/CSCS.2013.85

Ian J. Goodfellow, J. P.-A.-F. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems.

Karras, T. (2020). Analyzing and improving the image quality of StyleGAN. Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, (pp. 8110-8119).

Kosmyna, N. G. (2020). The thinking cap 2.0' preliminary study on fostering growth mindset of children by means of electroencephalography and perceived magic using artifacts from fictional sci-fi universes. Proc. Interaction Design and Children Conference, (pp. 458-469).

Mirsky, Y. &. (2021). The creation and detection of deepfakes: a survey. ACM Comput. Surveys.

O. L. Liu, B. B. (2012). Measuring learning outcomes in higher education: Motivation matters. Educational Researcher, 352-362.

Pataranutaporn, P., Danry, V., Leong, J., Punpongsanon, P., Novy, D., Maes, P., & Sra, M. (2021). AI-generated characters for supporting personalized learning and well-being. In Nature Machine Intelligence (Vol. 3, Issue 12). https://doi.org/10.1038/s42256-021-00417-9

Spinath, R. S. (2009). The importance of motivation as a predictor of school achievement. Learning and individual differences, 80-90.

Xu, Y., Vigil, V., Bustamante, A. S., & Warschauer, M. (2022). "Elinor's Talking to Me!":Integrating Conversational AI into Children's Narrative Science Programming. Conference on Human Factors in Computing Systems - Proceedings. https://doi.org/10.1145/3491102.3502050

Coqui GmbH. (2020). *Glow TTS - TTS 0.22.0 documentation*. https://docs.coqui.ai/en/dev/models/glow_tts.html

Dheeraj Jalali. (2020, June 9). *How Voice Computing is Building a More Accessible World | Voices | Voices*. https://www.voices.com/blog/text-to-speech-technology/

Dutoit, T. (1997). High-quality text-to-speech synthesis: An overview. *Journal Of Electrical And Electronics Engineering Australia*, *17*(1), 25–36.

Kim, J., Kim, S., Kong, J., & Yoon, S. (2020). Glow-tts: A generative flow for text-to-speech via monotonic alignment search. *Advances in Neural Information Processing Systems*, *33*, 8067–8077.

Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, & Dingzeyu Li. (2020). *MakeItTalk - SA'20*. https://people.umass.edu/~yangzhou/MakeItTalk/

Zhou, Y., Han, X., Shechtman, E., Echevarria, J., Kalogerakis, E., & Li, D. (2020). MakeItTalk: Speaker-Aware Talking-Head Animation. *ACM Transactions on Graphics*, *39*(6).