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

Learning with Virtual Avatars: Insights into Performance and Resource Needs

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

Abstract: This paper presents a novel method for generating speaking avatars from text, images, and custom audio, designed to enhance digital communication and virtual interaction. Method involves two primary components: image processing and audio synthesis. It starts with a static image of a person and applying facial recognition and animation algorithms to create a dynamic, lifelike avatar capable of realistic mouth movements and expressions. Concurrently, it utilizes text-to-speech (TTS) technology to convert written text into natural-sounding speech, tailored to match the avatar's identity and intended emotional tone. By integrating these two elements, it is ensured that the avatar's lip movements are synchronized with the generated audio, resulting in a seamless and engaging user experience. Additionally, by leveraging powerful CPUs and GPUs, it is demonstrated how current technology enables efficient and sophisticated video and audio generation. The findings underscore the need for optimized resource management to achieve balanced processing speed and output quality. As a potential direction of future research, it is proposed to test virtual avatars in real educational environments, with an emphasis on evaluating the effect on learning, motivation and engagement of students.

Keywords: sound generation, video generation, advanced hardware configurations

1. Introduction

In today's digital age, rapid technological advancements, particularly in artificial intelligence (AI), are transforming various industries and enhancing everyday life through automation, data analysis, and personalized services. One groundbreaking development is the ability to generate videos of people speaking using only text and photos, creating virtual characters that can revolutionize fields such as education, medicine, tourism, and marketing. In education, AI-driven virtual characters can personalize learning experiences, adapting to individual needs and learning styles, thus modernizing the school system. This concept is further supported by studies such as those by Hobert & Berens (2024), which highlight the role of digital tutors as intermediaries in educational settings, providing real-time, individualized support to students. This paper examines methods and techniques for sound and image generation, analyzes current models, and identifies the resources necessary for successful implementation.

2. Generation of Virtual Avatars

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

YourTTS is a state-of-the-art text-to-speech (TTS) system that represents a significant advancement in the field of speech synthesis. It utilizes a versatile and robust model designed to generate highly natural and expressive speech from text. Unlike traditional TTS systems, YourTTS can adapt to various accents and speaking styles, offering personalized and context-aware speech synthesis. This innovative approach not only enhances the quality of generated

speech but also ensures its applicability across different languages and dialects, making it a versatile tool for global applications (Casanova et al., 2021).

2.2 Video generation

MakeItTalk (Zhou et al., 2020) 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. It maps the audio features to corresponding facial movements, creating a seamless and natural-looking talking head.

Even though MakeItTalk produces somewhat realistic videos, its output is only 256x256 pixels. To address this limitation, CodeFormer (Zhou et al., 2022) is used for image enhancement. CodeFormer is a robust face restoration project that utilizes the Codebook Lookup Transformer to improve the quality of face images. By using CodeFormer, each frame of the generated video undergoes precise enhancement, ensuring that the result is clear, coherent, and visually striking.

2.3 System architecture

The video and sound generation system is built on a robust server architecture that is developed using Python's FastAPI framework, serves as the central hub for managing and processing complex tasks such as video generation and AI interaction. It leverages high-performance GPUs to ensure swift and effective video processing.

Figure 1 shows application screenshots of two main pages. Initially system requires the input of both an image and an audio file. Once these inputs are provided, user can either ask questions or provide text prompts. The server then processes these requests, using YourTTS for text-to-speech (TTS) synthesis to generate natural and expressive speech, and MakeItTalk for video generation. If user request high-quality video, the system further incorporates CodeFormer to enhance image quality.

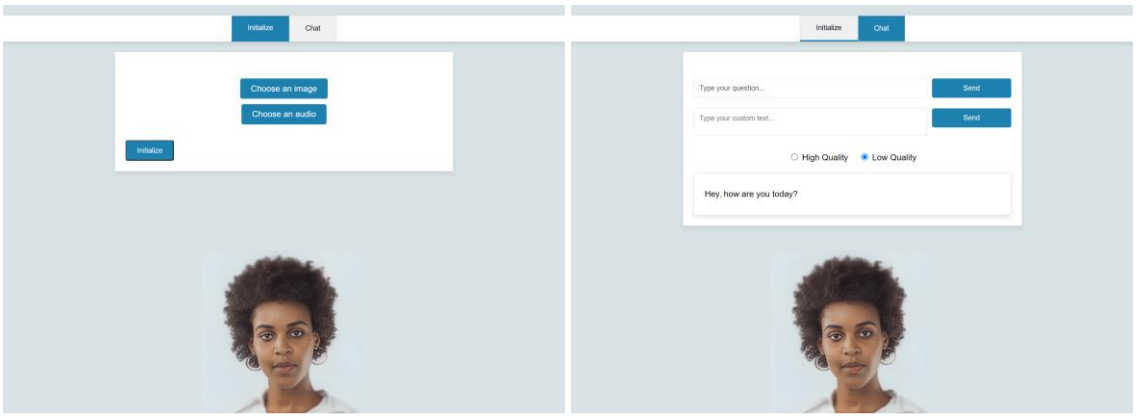

Figure 1. Application screenshots – showcase of initialization page and chat page

3. Results – performance testing

Performance evaluation is crucial for assessing system effectiveness. This section examines a setup with AMD EPYC 7763 processors and NVIDIA A100 graphics cards, focusing on their roles in handling various workloads. Key factors include image processing group size (number of images processed simultaneously), device (GPU usage status), quality level (whether video enhancement was performed) and processing time (time required for video generation). These factors provide insights into the performance capabilities and limitations of the system. Table 1 shows performance test results for entry text: "The serene lake reflected the colors of the sunset like a painter's masterpiece. In the bustling city, amidst the honking cars and hurried pedestrians, she found a moment of quiet in a quaint cafe.

Table 1. Performance test results with two sentences

Image processing batch size	Device	Quality level	Processing time (seconds)
16	CPU + GPU	Low	48
16	CPU + GPU	High	Out of memory
16	CPU	Low	140
8	CPU + GPU	Low	51
8	CPU + GPU	High	1.961
8	CPU	Low	103
4	CPU + GPU	Low	53
4	CPU + GPU	High	1.075
4	CPU	Low	123

In analyzing performance results for generating video from input text, key insights highlight the impact of hardware configurations. For a batch size of 16 images, the CPU + GPU setup processes low-quality outputs in 48 seconds but fails with high-quality outputs due to GPU memory limits. CPU-only processing for low-quality takes 140 seconds, showing the importance of GPU acceleration. For a batch size of 8, the CPU + GPU setup takes 51 seconds for low-quality and 1,961 seconds for high-quality outputs. The CPU-only configuration is faster than the batch size of 16 but slower than the combined setup, with 103 seconds for low-quality. For a batch size of 4, the CPU + GPU system handles low-quality outputs efficiently, while high-quality processing takes 1,075 seconds. CPU-only processing for this batch size takes 123 seconds, underscoring the GPU's performance benefits.

4. Conclusion

The performance of a video generation system using advanced hardware configurations highlights the importance of GPU acceleration, especially in educational settings where multimedia content can enhance learning experiences. Leveraging advanced CPUs and GPUs, the research demonstrates how optimized resource management enhances processing speed and output quality. By using images and voices, you can create virtual avatars that have the potential to enhance the learning experience. These advancements support educators in delivering personalized, dynamic and effective learning experiences, making complex concepts more accessible to students using virtual avatars. In conclusion, future research could explore using virtual avatars as learning assistants, such as avatars of historical figures, to enhance students' engagement, motivation, and learning in specific areas. Integrating these avatars into existing learning management systems could further support broader adoption and application in educational settings.

References

Casanova, E., Weber, J., Shulby, C., Junior, A. C., Gölge, E., & Ponti, M. A. (2021). YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for everyone. Proceedings of Machine Learning Research, 162, 2709–2720. https://arxiv.org/abs/2112.02418v4

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

Hobert, S., & Berens, F. (2024). Developing a digital tutor as an intermediary between students, teaching assistants, and lecturers. Educational Technology Research and Development, 72(2), 797–818. https://doi.org/10.1007/s11423-023-10293-2

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.

Zhou, S., Chan, K. C. K., Li, C., & Loy, C. C. (2022). Towards Robust Blind Face Restoration with Codebook Lookup TransFormer. NeurIPS.

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).