Pacific Graphics (2025)
P.-H. Han and H.-R. Tsai (Editors)

Avatar Animations and Audio Fillers for Managing Response Delays

Gopi Krishnan Singaravelan[1] Zhi Lynn Lay[1] Ping-Hsuan Han[1]

[1]National Taipei University of Technology

Abstract

This study presents techniques for managing response delays in avatars with large language models (LLMs) to enhance user interaction. While existing avatar-based LLMs focus on human-like conversational abilities, they often overlook the impact of response delays on user experience. Our system strategically reframes these delays as opportunities to enhance the perceived humanness of the avatar by incorporating emotion-based animations, a companion pet, and contextually appropriate audio fillers. Through thoughtful audio-visual design and user interface enhancements during waiting periods, the demo showcases how effective delay management can sustain engagement, foster natural interactions, and turn waiting moments into meaningful elements of the conversational experience.

CCS Concepts

• ***Methods and Applications*** → *Artificial Intelligence ;* • ***Interaction*** → *User Interface Design;* • ***Animation*** → *Behavioral Animation;*

1. Introduction and Motivation

With advancements in large language models (LLMs) like OpenAI, Gemini. chatbots now generate accurate and nuanced responses [JRB24]. While text-based chatbots are common, avatar-based systems take the next step in interaction by integrating voice, expressions, and gestures for more engaging conversations [YMY*23]. Prior work shows that visual, identity, and conversational cues shape humanness perceptions [GS19], and recent studies highlight that anthropomorphic design cues appearance and conversational style jointly influence social presence, trust, satisfaction, and visual attention. [CGR*24]

At the same time, delay management is critical: conversational and gestural fillers can reduce perceived latency and support natural dialogue flow with digital humans. [KL22] These insights underscore the role of anthropomorphic cues and latency-handling strategies in shaping user experience.

In this demo, we present an avatar-based chatbot system with a primary avatar, companion avatar, and contextually appropriate audio fillers. Leveraging TTS, STT, and LLM technologies, the system manages response delays through expressive animations, companion interactions, and audio Fillers, sustaining engagement and enhancing conversational flow.

2. System and Implementation

We present the Enhanced Avatar Chatbot System, an emotionally expressive, avatar-based communication platform designed to enable natural human–AI interaction through spoken language, rich visual feedback, and adaptive emotional expression. The system incorporates a complete conversational pipeline: Speech-to-Text (STT) using *Whisper*, Text-to-Speech (TTS) using *ElevenLabs*, and language generation via the *LLaMA 3.2* model from Ollama implemented within *Unity 2022.3.22f1*.

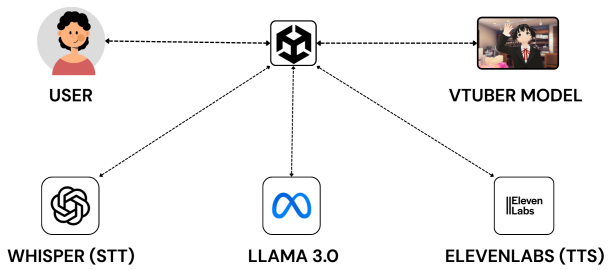

Figure 1: *System Overview.*

Figure 1 illustrates the System overview of the avatar-based chatbot system. The user's speech is first captured in Unity and transcribed into text using Whisper (STT). This text is then processed by LLaMA 3.2 to generate an appropriate response. The output is converted into natural-sounding speech through ElevenLabs (TTS). Finally, Unity synchronizes the generated audio with the VTuber model, delivering real-time facial and body animations that create an engaging and immersive conversational experience.

DOI: 10.2312/pg.20251316

To enable adaptive responses, the system uses a **prompt-engineered LLaMA-based emotion recognition module**. After the LLM generates a response, the text is passed to a specially designed LLaMA prompt that performs two tasks:

1. **Emotion Classification** – Identifies one of five target emotional states: joy, anger, sadness, fear, or neutral.
2. **Intensity Analysis** – Determines the strength of the detected emotion, categorizing it as low, moderate, or strong.

The detected emotion and intensity directly drive the avatar's behavior. For example, high-intensity joy triggers wide smile blendshapes, lively eye animations, brighter facial textures, celebratory particle effects, and energetic full-body motions. Conversely, low-intensity sadness might produce subtle downward gaze, dimmer facial tones, and slower body movement.

In addition, the system employs **delay handling mechanisms** audio fillers, emotion-based animations, and a companion pet avatar to sustain engagement, reduce perceived waiting time, and make interactions feel more natural and immersive.

3. Application

A user engaging with the Enhanced Avatar Chatbot System, where real-time speech and expressive animations bring conversations to life in a virtual café environment.

Figure 2: *The Enhanced Avatar Chatbot.*

The Enhanced Avatar Chatbot System incorporates a delay management strategy that continuously monitors the elapsed time between a user's input and the LLM's generated response. Specific features are triggered at predefined intervals to maintain engagement and reduce perceived waiting time.

Table 1: *Delay Intervals and Behaviors*

Delay	Behavior
$\geq$ 1s	Friendly filler ("Hmm…", "Let me think…").
$\geq$ 3s	Emotion-based filler (previous emotion).
$\geq$ 5s	Standard filler ("Still thinking…").
$\geq$ 7s	Companion pet animation (loop).

Table 1 If a delay occurs, the system activates features in sequence. After 1 second, it plays a friendly filler such as "Hmm…" or "Ahan" to mimic natural human like sounds. If the delay exceeds 3 seconds, it triggers an emotion-based filler that aligns with the user's previous emotional state. At 5 seconds, a standard filler such as "Still processing…" reassures the user that the system is working. For delays beyond this point, a companion pet avatar plays looping animations to visually engage and distract the user until the response is ready.

Beyond delay handling, the system supports easy scenario customization through prompt engineering, enabling developers to switch between conversational roles or environments such as a café barista, travel guide, or virtual tutor without modifying the core architecture.Figure 2 illustrates both the delay-handling mechanism and the café scenario in action. Additionally, the system can remember user-specific details, such as their name, to personalize interactions and foster a sense of familiarity over time.

This combination of tiered delay handling, flexible scenario adaptation, and personalized memory delivers a more natural, immersive, and user-centric conversational experience.

4. Discussion and limitation

The Enhanced Avatar Chatbot System shows that response delays can be reframed into engaging moments through tiered delay handling. Friendly fillers at short delays, emotion-based fillers at medium delays, and companion pet animations at longer delays helped maintain user attention and reduce perceived waiting time. Emotion-aligned fillers and basic personalization made the system adaptable to multiple domains without architectural changes. However, a key limitation is that scenarios are generated only through pre-trained data sets or prompt design. The system is not capable of fully open-ended conversation or understanding unscripted scenarios. Overall, combining audio, visual, and emotional feedback created a more immersive and human-like conversational experience.

5. Conclusion and Future work

The Enhanced Avatar Chatbot System turns response delays into engaging moments through audio fillers, emotion-based animations, and companion pet interactions. Future work will explore multimodal interactions such as gesture recognition, gaze detection, and advanced emotion analysis to further enhance engagement during delays.

References

[CGR*24] CHEN J., GUO F., REN Z., LI M., HAM J.: Effects of anthropomorphic design cues of chatbots on users' perception and visual behaviors. *International journal of human–computer interaction 40*, 14 (2024), 3636–3654. 1

[GS19] GO E., SUNDAR S. S.: Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in human behavior 97* (2019), 304–316. 1

[JRB24] JOHN K. S., ROY G. A., BINDHYA P.: Llm based 3d avatar assistant. In *2024 1st International Conference on Trends in Engineering Systems and Technologies (ICTEST)* (2024), IEEE, pp. 1–5. 1

[KL22] KUM J., LEE M.: Can gestural filler reduce user-perceived latency in conversation with digital humans? *Applied Sciences 12*, 21 (2022), 10972. 1

[YMY*23] YAMAZAKI T., MIZUMOTO T., YOSHIKAWA K., OHAGI M., KAWAMOTO T., SATO T.: An open-domain avatar chatbot by exploiting a large language model. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (Prague, Czechia, Sept. 2023), Stoyanchev S., Joty S., Schlangen D., Dusek O., Kennington C., Alikhani M., (Eds.), Association for Computational Linguistics, pp. 428–432. URL: https://aclanthology.org/2023.sigdial-1.40, doi:10.18653/v1/2023.sigdial-1.40. 1