AVIN-Chat: An Audio-Visual Interactive Chatbot System with Emotional State Tuning

Chanhyuk Park , **Jungbin Cho** , **Junwan Kim** , **Seongmin Lee** , **Jungsu Kim**
and **Sanghoon Lee**

Yonsei University

{hanul500, whwjdqls99, jw1510, lseong721, integer528, slee}@yonsei.ac.kr,

Abstract

This work presents an audio-visual interactive chatbot (AVIN-Chat) system that allows users to have face-to-face conversations with 3D avatars in real-time. Compared to the previous chatbot services, which provide text-only or speech-only communications, the proposed AVIN-Chat can offer audio-visual communications providing users with a superior experience quality. In addition, the proposed AVIN-Chat emotionally speaks and expresses according to the user's emotional state. Thus, it enables users to establish a strong bond with the chatbot system, increasing the user's immersion. Through user subjective tests, it is demonstrated that the proposed system provides users with a higher sense of immersion than previous chatbot systems. The demonstration video is available at https://www.youtube.com/watch?v=Z74uIV9k7_k.

1 Introduction

With the recent advancements of large language models (LLMs) [Brown *et al.*, 2020], interactive AI chatbot services have gained attention from various industries including customer support, virtual assistants, and education. These services provide users with great convenience but their text-only[1] or audio-only[2] interface hinders immersive user experience.

To increase user immersion in these chatbot services, as shown in Fig. 1, we present an audio-visual interactive AI chatbot system, which listens and responds to users with a realistic and emotional talking face and voice. To build this system, we composed our system with three distinctive sub-modules: (1) construct a deformable realistic face, (2) listen and talk to users in real-time, and (3) generate the avatar's facial expressions and lip movement in sync with its response. Additionally, we enable our system with emotional expressions so that it can adapt to users' feedback during conversations providing an immersive experience.

Current realistic face reconstruction methods are capable of generating realistic facial mesh along with high-quality

---

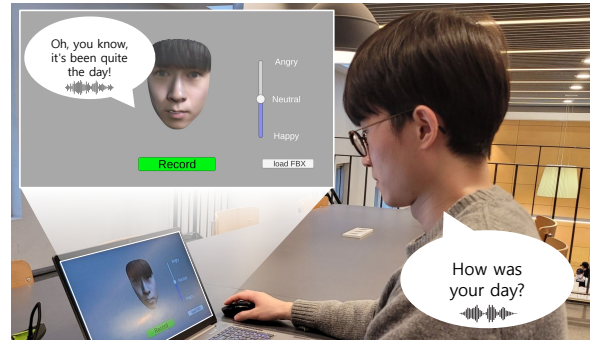[1]https://juji.io/

[2]https://assistant.google.com/

Figure 1: Example of the AVIN-Chat use. AVIN-Chat receives the user's speech and generates an audio-visual response in real-time.

textures [Zheng *et al.*, 2023; Zielonka *et al.*, 2022; Kang *et al.*, 2021; Kang *et al.*, 2020; Lee *et al.*, 2023a] but require expensive computational costs for animation. Thus, these methods cannot be applied to the real-time chatbot system. Note that as we focus on animating arbitrary facial meshes, we use a template-free 3D face instead of a template-based 3D face model. To enable template-free real-time 3D face animation, we generate facial blendshapes for expression animation from the reconstructed facial mesh. Once blendshapes are defined, it is possible to express facial expressions with only a single linear calculation, so 3D faces can be animated with no latency [Lee *et al.*, 2023c; Lee *et al.*, 2023b].

For a speech-driven interactive chatbot system, we employ the speech-to-text (STT) and text-to-speech (TTS) models. The user's speech is converted into text using the STT model and then fed to the ChatGPT[3], which is the state-of-the-art LLM. ChatGPT acts like a brain in listening to users and generating answers in the proposed system. Thus, after the user's statement is fed, ChatGPT generates answers and speaks using the TTS model. Here, we perform in-context learning on ChatGPT to generate the emotional response according to the user-defined emotions. For example, if the user wants to communicate with happy emotions, the user can adjust the ChatGPT to generate a response with happy emotions. Finally, to animate 3D faces synchronize with the ChatGPT's response, we employ EmoTalk, which is the speech-driven

---
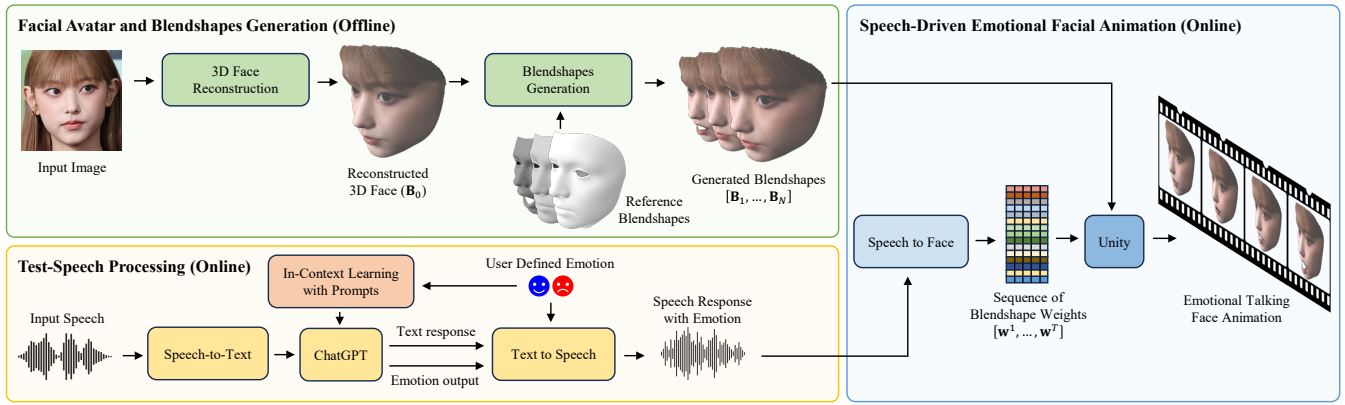
[3]https://openai.com/blog/chatgpt

Figure 2: Overall pipeline of the proposed audio-visual interactive chatbot (AVIN-Chat) system. AVIN-Chat is constructed with three sub-modules: 1) facial avatar and blendshapes generation, 2) text-speech processing, and 3) speech-driven emotional facial animation modules.

3D face animation model [Peng *et al.*, 2023]. EmoTalk generates facial expressions and lip movements from the speech input by estimating the coefficient of the facial blendshapes. As we already built the facial blend shape that aligns with those of EmoTalk, facial expressions and lip movements can be efficiently animated with no latency.

Therefore, we present an audio-visual interactive chatbot (AVIN-Chat) system that allows users to communicate with one avatar who has the face and voice they want. Through the subjective test, we evaluate whether the proposed audio-visual chatbot system enables intimate communication compared to other text-only or audio-only AI chatbot systems.

2 Audio-Visual Interactive Chatbot

2.1 Overall Pipeline

Figure 2 illustrates the overall pipeline of the proposed AVIN-Chat system. The proposed system is composed of facial avatar generation, text-speech interaction, and speech-driven facial animation. Subsequent sections will provide detailed descriptions of each module.

Facial Avatar and Blendshapes Generation (Offline)

The goal of facial avatar generation is to reconstruct a realistic 3D facial mesh with high-quality texture and then generate its blendshapes for efficient facial animation. We first reconstruct a realistic facial mesh and then generate the 52 standard blendshapes allowing it to be deformable. For the facial reconstruction, we adopt the HRN from [Lei *et al.*, 2023]. HRN decouples facial geometry into (1) the low-frequency part for a coarse shape of the overall facial mesh, (2) the mid-frequency for local shapes, and (3) the high-frequency part for detailed information such as wrinkles. This hierarchical fashion allows us to reconstruct a photo-realistic facial mesh from a single image. Note that although we used a template-based 3D face reconstruction approach, we assume that the reconstructed face is a non-template face mesh for system scalability.

To make the reconstructed non-template facial mesh deformable, we generate blendshapes based on the 52 facial ac-

tion units[4]. For the blendshapes generation, we employ the LDT [Onizuka *et al.*, 2019] that deforms the facial mesh according to the target expressions using landmarks as guidance. Here, since LDT requires facial landmarks for facial deformation, we extract landmarks from HRN reconstructed faces. We extract 52 facial landmarks in our implementation. If we use the face of a character who is difficult to automatically extract landmarks, we should manually designate the landmarks.

Text-Speech Processing (Online)

Text-speech interaction is to listen user's speech and generate a response. To generate the response we employ the OpenAI's ChatGPT, which is the state-of-the-art LLM. To feed the user's speech to the ChatGPT, we first convert speech to text using the state-of-the-art STT model named Whisper [Radford *et al.*, 2023]. Whisper exhibits remarkable robustness in dealing with different audio conditions, effectively managing challenges such as background noise and various speech accents. From the encoded text, ChatGPT generates the text response. Here, to generate natural-sounding responses, we perform in-context learning on ChatGPT using prompts.

Since the ChatGPT generates text responses, we adopt the TTS model to convert text to speech. For TTS, we utilize EmotiVoice[5], an improved version of PromptTTS [Guo *et al.*, 2023]. EmotiVoice allows us to adjust the emotion in the voice by providing certain text prompts. The details for the ChatGPT and EmotiVoice prompts are described in Sec. 2.2.

Speech-Driven Emotional Facial Animation (Online)

We employ the speech-driven facial animation model to generate the emotional talking faces from the ChatGPT's emotional speech response. For the emotional talking face animation, we utilize the EmoTalk [Peng *et al.*, 2023]. EmoTalk estimates a sequence of blendshapes weights $[\mathbf{w}^1, \ldots \mathbf{w}^T]$ where $\mathbf{w} = [w_1^T, \ldots, w_N^T]$ is the set of the N-blendshape and T is the number of frames. Thus, the facial shape of tth frame $\mathbf{S}^t$ is computed as $\mathbf{S}^t = \mathbf{B}_0 + \mathbf{w}^t(\mathbf{B} - \mathbf{B}_0)$, where $\mathbf{B}_0$ is the

---

[4]https://arkit-face-blendshapes.com

[5]https://github.com/netease-youdao/EmotiVoice

reconstructed neutral face and $\mathbf{B} = [\mathbf{B}_1, \ldots, \mathbf{B}_N]$ is the generated N-blendshapes. Generated face sequences $[\mathbf{S}^1, \ldots, \mathbf{S}^T]$ are visualized using Unity.

2.2 In-Context Learning with Prompts

It is shown that prompts significantly influence the output of LLMs [Schick and Schütze, 2021]. Therefore, there have been many attempts to control the results of LLMs through various prompting techniques [Shin *et al.*, 2020; Lester *et al.*, 2021]. We utilized multiple prompts to build a chatbot system capable of natural conversation. In the proposed system, since ChatGPT preserves the context of the conversation, we set various conditions through prompts before starting the conversation to maintain consistency in the subsequent dialogue. To give the user a more human-like experience, we ensured the responses were not too long, avoided repetition of the same phrases, and defined do's and don'ts for responding in a friendly manner. Additionally, we informed ChatGPT that the response text would be converted to speech, instructing it not to generate responses that are hard to verbalize, like emojis.

Furthermore, we design our system so that users can input their desired emotional state. Using the user interface in Unity, the user can easily select the desired emotional state and the selected emotion is fed to the ChatGPT and EmotiVoice as prompts. If users are not satisfied with the chatbot's current emotional state, they can simply adjust the chatbot's emotional state during the conversation using the graphical user interface (GUI) in Unity. By implementing this approach, we observed that the generated responses were significantly more cheerful and friend-like compared to previous answers. It implies that the proposed chatbot is actively engaging in leading the conversation.

2.3 Implementation Details

The overall structure of the AVIN-Chat consists of a Python backend server and a Unity frontend. In the facial avatar generation module, Unity sends an image captured by a webcam to the server. Then using the HRN, the server generates a 3D facial mesh in OBJ format along with a texture file. Using the OBJ file as a base, the LDT model creates 52 additional OBJ files, and the Blender Python API combines them to create an FBX file, which is the final blendshapes. Unity receives this 3D FBX with its texture file and displays the constructed face to the user. In the conversational part, Unity records the user's voice, sending it to the server. In return, Unity receives an answering voice along with a list of blendshapes parameters. Once this exchange is done, Unity runs the audio and animates the facial mesh with the blendshapes parameters. For the emotional tuning, we've configured Unity to communicate with a backend API to change ChatGPT's prompts to match the user-defined emotions. Details of our system's graphical user interface are shown in Fig. 3. We used AMD Ryzen 7 5800u and AMD Radeon (TM) Graphics for the client and AMD Ryzen 5 5600X and NVIDIA RTX 3060 Ti for the server.
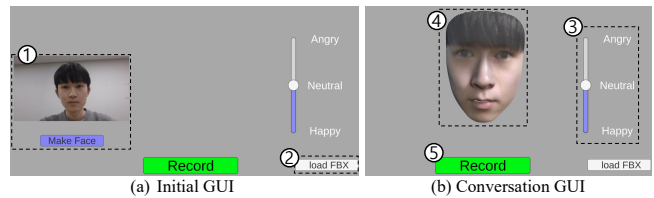

(a) Initial GUI   (b) Conversation GUI

Figure 3: GUI visualization and example of actual use of the AVIN-Chat: (a) Initial GUI where users can ① capture an image or ② load an FBX file, and (b) conversation GUI where users can ③ define emotions and talk to ④ facial avatar by using the ⑤ record button.

3 Experimental Results

We evaluate our proposed system on what conversations users prefer with different AI chatbots, such as text-only and speech-only systems, compared to ours. The text-only system utilizes ChatGPT which interacts with the user only through text inputs and outputs. The speech-only system is composed of ChatGPT, Whisper, and EmotiVoice. Note that the speech-only system also can generate emotional answers because it utilizes EmotiVoice. For fair comparisons, we use the same model weights as those used in the proposed AVIN-Chat. We recruited 11 participants to score each system's preference over the proposed AVIN-Chat. When scoring, participants are asked to measure their preferences in terms of intimacy, immersiveness, empathy, and overall satisfaction. Figure 4 shows the subjective test results. It shows that the text-only chatbot achieves better performance than the speech-only chatbot. Participants reported that visual stimulus, although being only text, enhanced user experience more than voice stimulus. In addition, the results demonstrate that the proposed AVIN-Chat significantly outperforms other chatbot systems in terms of preference by providing both visual and auditory stimulus.
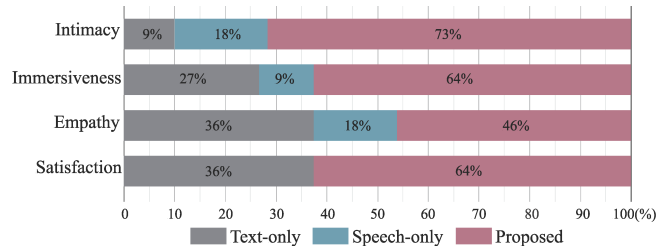

Figure 4: User preference scores for different chatbot systems on intimacy, immersiveness, empathy, and overall satisfaction.

4 Conclusion

In this paper, we have presented an audio-visual interactive chatbot, an end-to-end system for having face-to-face conversations with facial avatars reconstructed from a single image. Experimental results demonstrate that the proposed system can provide more immersive communication compared to text-only and speech-only chatbot systems. We believe that the proposed system can encourage users to use chatbot services and can be applied to various applications such as education and medical therapy. Our future work is to incorporate body gestures into a talking face for more realistic virtual communications.

Acknowledgements

Contribution Statement

Chanhyuk Park, Jungbin Cho, and Junwan Kim equally contributed to this work. Sanghoon Lee is the corresponding author of this work.

References

[Brown *et al.*, 2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[Guo *et al.*, 2023] Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xu Tan. Promptts: Controllable text-to-speech with text descriptions. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[Kang *et al.*, 2020] Jiwoo Kang, Seongmin Lee, and Sanghoon Lee. Uv completion with self-referenced discrimination. In *Eurographics (Short Papers)*, pages 61–64, 2020.

[Kang *et al.*, 2021] Jiwoo Kang, Seongmin Lee, and Sanghoon Lee. Competitive learning of facial fitting and synthesis using uv energy. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 52(5):2858–2873, 2021.

[Lee *et al.*, 2023a] Kyungjune Lee, Jeonghaeng Lee, Hyucksang Lee, Mingyu Jang, Seongmin Lee, and Sanghoon Lee. Faceclone: Interactive facial shape and motion cloning system using multi-view images. In *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 512–513. IEEE, 2023.

[Lee *et al.*, 2023b] Seongmin Lee, Hyunse Yoon, Jiwoo Kang, Jungsu Kim, Jiwan Son, Jungwoo Huh, and Sanghoon Lee. Video-based stabilized 3d face alignment using temporal multi-discrimination. In *2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6. IEEE, 2023.

[Lee *et al.*, 2023c] Seongmin Lee, Hyunse Yoon, Sohyun Park, Sanghoon Lee, and Jiwoo Kang. Stabilized temporal 3d face alignment using landmark displacement learning. *Electronics*, 12(17):3735, 2023.

[Lei *et al.*, 2023] Biwen Lei, Jianqiang Ren, Mengyang Feng, Miaomiao Cui, and Xuansong Xie. A hierarchical representation network for accurate and detailed face reconstruction from in-the-wild images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 394–403, 2023.

[Lester *et al.*, 2021] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, 2021.

[Onizuka *et al.*, 2019] Hayato Onizuka, Diego Thomas, Hideaki Uchiyama, and Rin-ichiro Taniguchi. Landmark-guided deformation transfer of template facial expressions for automatic generation of avatar blendshapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[Peng *et al.*, 2023] Ziqiao Peng, Haoyu Wu, Zhenbo Song, Hao Xu, Xiangyu Zhu, Jun He, Hongyan Liu, and Zhaoxin Fan. Emotalk: Speech-driven emotional disentanglement for 3d face animation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20687–20697, 2023.

[Radford *et al.*, 2023] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023.

[Schick and Schütze, 2021] Timo Schick and Hinrich Schütze. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, 2021.

[Shin *et al.*, 2020] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, 2020.

[Zheng *et al.*, 2023] Yufeng Zheng, Wang Yifan, Gordon Wetzstein, Michael J. Black, and Otmar Hilliges. Pointavatar: Deformable point-based head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[Zielonka *et al.*, 2022] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Instant volumetric head avatars. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4574–4584, 2022.