Exploring Korean AI Companion Possibilities with Live2D and Empathic Voice Interaction

Jaeyoung Suh, Mingyu Jeon

tjwodud04@gmail.com, jkmcoma7@gmail.com

Abstract

This study explores the growing trend of fostering emotional connections with AI through natural voice interactions. It presents a web-based Korean voice interaction system that integrates high-quality speech synthesis with a Live2D character. Research suggests that voice-based communication enhances emotional engagement more effectively than text-based methods, particularly in gaming and virtual reality. The system leverages OpenAI's gpt-4o-audio-preview model to generate expressive speech while synchronized with visual elements. Designed to enhance the synergy between auditory and visual channels, it delivers a well-balanced multimodal experience. Quantitative and qualitative evaluations confirm enhanced user immersion and emotional satisfaction, supporting deeper human-AI interactions. This study provides a practical framework for advancing multimodal AI interactions, offering valuable insights for both research and real-world applications in immersive AI communication.

Index Terms: Live2D Animation, Korean Conversational AI, Empathetic AI Interaction

1. Introduction

Recent studies indicate that users are increasingly experiencing a growing sense of connection with AI through natural, empathetic dialogue [1]. Research demonstrates that spoken interactions foster stronger emotional bonds than text-based communication, thereby enhancing user immersion in digital environments [2]. In role-playing games and interactive applications, AI-driven non-player characters (NPCs) have been reported to deepen engagement and contribute to more memorable and meaningful experiences [3]. These findings highlight the potential of natural speech to bridge the gap between human and AI interactions.

Conversations with AI companions are becoming an integral part of daily social and emotional life [4]. Technological advancements, such as OpenAI's GPT-4o model [5], which enhances the naturalness and emotional expressiveness of synthesized speech, demonstrate AI's potential for lifelike interactions. Additionally, research suggests that incorporating visual characters in virtual reality (VR) and mixed-reality (MR) environments enhances emotional engagement and immersion [6]. Similarly, auditory manipulations, such as spatial sound design and mixed-reality awareness cues, improve user perception and foster stronger emotional connections in AI-driven interactions [7].

Building upon these advancements, this study presents a web-based system for Korean language voice interactions that integrates high-quality natural speech with dynamic Live2D character representations [8]. By combining state-of-the-art voice synthesis with visually engaging avatars, the system delivers a balanced multimodal experience optimized for both auditory and visual sensory engagement. The system utilizes the gpt-4o-audio-preview model [9], known for its real-time emotional expressiveness.

This research addresses the challenge of achieving seamless multimodal AI interactions. Specifically, it aims to provide users with a more immersive and emotionally engaging experience, potentially contributing to healthier human–AI relationships. Additionally, this study offers a framework for integrating advanced voice synthesis with dynamic visual representations, paving the way for future developments in interactive AI systems.

This study first introduces the methodology behind the proposed approach, explaining how the system's structure and interaction flow are designed to enhance user immersion. It then examines the system's effectiveness through both quantitative and qualitative evaluations. Based on these findings, we discuss the limitations of the current approach and explore potential directions for future research. Finally, we summarize the key contributions and implications of this study.

2. Method

2.1. Voice Interaction using gpt-4o-audio-preview Model

The system utilizes OpenAI's gpt-4o-audio-preview model to generate natural, emotionally expressive speech [9]. This model supports robust multimodal processing, enabling real-time, context-aware vocal responses through both text-to-audio and audio-to-audio conversions. By incorporating explicit emotion-guiding instructions, intonation and prosody are fine-tuned to convey the intended sentiment, making interactions sound warm and authentic.

Designed for adaptability, the model generates vocal responses suitable for diverse applications, ranging from interactive gaming to digital assistants. To ensure low latency and high-fidelity synthesis, OpenAI's API is leveraged, minimizing delays while preserving sound quality. Consequently, this framework enhances user engagement while fostering more natural human-AI interactions.

The approach prioritizes simplicity, clarity, and emotional accuracy, delivering a technically robust yet user-friendly conversational experience. Expressiveness and realism are carefully balanced to strengthen the emotional connection between AI and users.

2.2. Enhancing Emotion and Speech Recognition through Chain-of-Thought Prompting

Building upon the strong foundation of voice synthesis, the Chain-of-Thought (CoT) prompting technique is employed to improve emotion and speech recognition [10]. By generating intermediate reasoning steps, the model enhances its understanding of speech content, contextual nuances, and speaker intent.

A key challenge in speech recognition, particularly for less widely spoken languages such as Korean, is transcription accuracy. Many multilingual large language models (MLLMs) and speech-to-text (STT) systems, including the gpt-4o-audio-preview model, often exhibit reduced performance for these languages compared to English [11], leading to difficulties in accurately interpreting user inputs.

Inspired by human strategies for compensating incomplete or unclear speech using contextual cues and prior knowledge [12], CoT prompting is applied to iteratively refine speech recognition outputs. Initially, the user's speech is transcribed into text via an STT process. The model then analyzes the transcribed text to infer contextual cues, subject matter, and speaker intent. Through iterative refinement, comprehension improves, resulting in more accurate and contextually appropriate interpretations.

Figure 1 visually outlines the iterative refinement process enabled by CoT prompting, thereby demonstrating its role in enhancing overall speech comprehension.

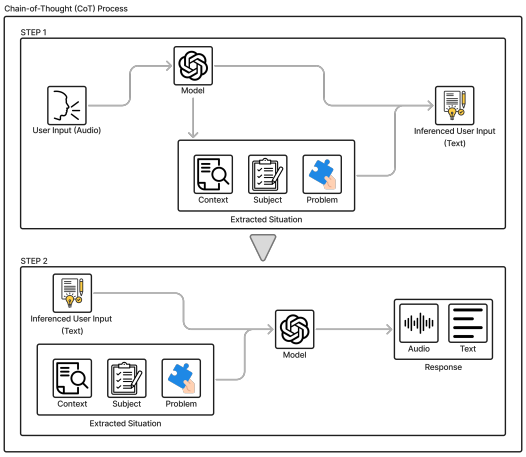

Figure 1: *Schematic representation of the CoT-based prompting flow*

2.3. Visual Representation with Live2D Character Kei

Parallel to the audio components, a Live2D character, Kei [13], is integrated to provide a visually engaging representation synchronized with the generated speech. Live2D technology transforms static illustrations into animated characters through a parameterized model, supporting real-time lip-syncing to ensure natural alignment between speech output and character motion.

Kei is equipped with predefined motion configurations and animation files to enhance expressiveness, enabling smooth and realistic visual feedback. Real-time rendering and animation control are achieved using the Live2D Cubism SDK, ensuring immediate responses to user interactions. Thus, the integration of dynamic voice synthesis with interactive visual elements reinforces AI presence and enhances user engagement.

3. Architecture and Interaction Process

3.1. Graphical User Interface Layout

The graphical user interface (GUI) is designed to support intuitive and immersive interactions. As illustrated in Figure 2, the interface integrates visual and auditory elements to create a cohesive user experience.

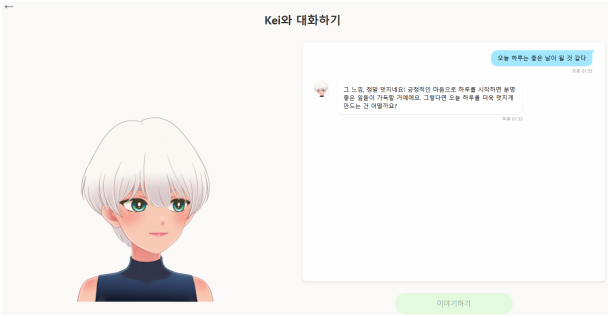

Figure 2: *Example interaction with Live2D Character Kei*

The primary features include components like below.

- **Character Display:** A Live2D-rendered AI character that dynamically reacts to user interactions with synchronized lip movements and facial expressions.
- **Speech Interaction Mechanism:** A dedicated voice input button that enables users to initiate interactions, with spoken inputs processed in real time through the gpt-4o-audio-preview model.
- **Conversational Flow:** A chat history window displaying transcribed text for both user inputs and AI-generated responses, ensuring accessibility and continuity.

This integrated design not only reinforces the naturalness of conversations by aligning visual cues with vocal outputs but also fosters a balanced multimodal experience. In other words, the seamless combination of these components deepens user engagement and strengthens the emotional resonance of AI-driven dialogues.

The voice interaction process is structured to maximize the strengths of both speech recognition and synthesis, as outlined in Figure 3.

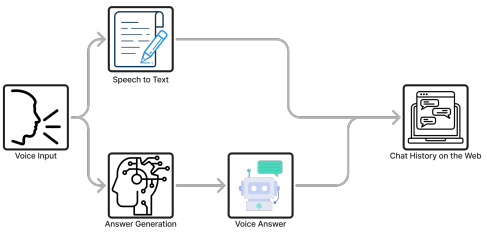

Figure 3: *The overall flow of interaction*

When a user initiates a voice input, the system processes it simultaneously through two parallel pathways:

Text Conversion Pathway: The first pathway employs the whisper-1 model to convert spoken input into text. This transcribed text is immediately stored in the conversation history, which ensures contextual continuity and provides a reliable backup in case of network delays or recognition errors. Such persistent text records are especially beneficial for applications like role-playing games and interactive simulations, where narrative consistency is crucial.

Voice Synthesis Pathway: Concurrently, the same voice input is routed through the gpt-4o-audio-preview model. This model analyzes the input's context and emotional nuances to generate natural and context-aware vocal responses. The synthesized speech is then played back in real time, while the system synchronizes the lip movements of the Live2D character with the audio output, resulting in a cohesive blend of visual and auditory feedback.

By employing this dual-processing architecture, the system not only mitigates challenges such as network latency and transcription errors but also creates a seamless interaction flow. In summary, the combination of immediate text backup and real-time, emotionally expressive speech synthesis significantly enhances human-AI communication, paving the way for more immersive and engaging applications.

4. Audio Model Evaluation

4.1. Speech-to-Text Performance

Speech recognition performance was evaluated by comparing a baseline STT model and a CoT-enhanced model for both English and Korean. For English, the LibriSpeech ASR dataset [14] was used, while the AIHub Dialogue Speech Dataset [15] was employed for Korean.

Transcription quality was assessed using ROUGE-1, ROUGE-2, ROUGE-L, and BLEU scores. ROUGE-N measures the n-gram overlap between reference and candidate transcriptions, emphasizing recall:

$$\text{ROUGE-N} = \frac{\sum_{S \in \{\text{Reference Summaries}\}} \sum_{\text{gram}_n \in S} \text{Count}_{\text{match}}(\text{gram}_n)}{\sum_{S \in \{\text{Reference Summaries}\}} \sum_{\text{gram}_n \in S} \text{Count}(\text{gram}_n)} \tag{1}$$

BLEU evaluates n-gram precision while applying a brevity penalty to discourage overly short outputs:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{2}$$

where BP is the brevity penalty, w_n represents the weight for n-gram precision p_n, and N is typically set to 4.

As shown in Table 1, both models perform well in English; however, Korean STT results are lower due to linguistic complexity and limited annotated data. Notably, the CoT model significantly improves Korean transcription—especially in ROUGE-2 and BLEU scores—indicating its effectiveness in capturing contextual relationships and resolving phoneme ambiguities. Consequently, this improvement lays a strong foundation for further emotion analysis.

4.2. Multi-Label Emotion Classification

For multi-label emotion classification, the Korean AIHub Dialogue Speech Dataset was used, annotated with seven emotions: happiness, anger, disgust, fear, neutral, sadness, and surprise. Evaluation metrics included Hamming Loss, Micro F1 Score, Macro F1 Score, and Jaccard Similarity.

Table 1: *STT Model Performance Comparison*

Model	ROUGE-1	ROUGE-2	ROUGE-L	BLEU
Eng Base	0.9722	0.9538	0.9722	0.9325
Eng CoT	0.9631	0.9369	0.9631	0.9117
Ko Base	0.7319	0.6278	0.7319	0.5263
Ko CoT	**0.7912**	**0.6728**	**0.7912**	**0.5642**

Hamming Loss quantifies the proportion of misclassified labels:

$$\text{Hamming Loss} = \frac{1}{N \times L} \sum_{i=1}^{N} \sum_{j=1}^{L} \mathbb{I}(y_{ij} \neq \hat{y}_{ij}) \tag{3}$$

The F1 Score, representing the harmonic mean of precision and recall, is computed as:

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

Micro F1 Score aggregates contributions across all classes, while Macro F1 Score averages the F1 Scores of individual classes. Jaccard Similarity evaluates the intersection-over-union of predicted and true labels:

$$\text{Jaccard Similarity} = \frac{|Y_{\text{true}} \cap Y_{\text{pred}}|}{|Y_{\text{true}} \cup Y_{\text{pred}}|} \tag{5}$$

As detailed in Table 2, the CoT model consistently outperforms the baseline across all metrics, demonstrating its superior ability to capture nuanced emotional expressions.

Table 2: *Emotion Classification Model Performance Comparison*

Model	Hamming Loss (↓)	Micro F1 Score (↑)	Macro F1 Score (↑)	Jaccard Similarity (↑)
Base Model	0.3086	0.4545	0.4665	0.3830
CoT Model	**0.2500**	**0.5698**	**0.5853**	**0.4697**

4.3. Qualitative Evaluation

To complement the quantitative metrics, a qualitative evaluation was conducted using a 1–100 Likert scale across five dimensions:

- **Q1: Emotion Detection Capability** - Accuracy in detecting user emotions
- **Q2: Naturalness of Korean Sentence Structure** - Fluency and grammatical correctness
- **Q3: Naturalness of the Voice** - Realism and pleasantness of the synthesized speech
- **Q4: Variety of Character Facial Expressions** – Diversity and appropriateness of visual expressions
- **Q5: Overall Experience** - General satisfaction with the interaction

Figure 4 shows that the CoT-applied model outperforms the Base model in most categories, particularly in **Q1** (Emotion Detection Capability) and **Q5** (Overall Experience). However, for **Q3** (Naturalness of the Voice), the Base model received a slightly higher satisfaction score. One possible explanation is

that the CoT model tends to generate longer or more complex Korean utterances, which can inadvertently highlight subtle unnatural phrasing or intonation issues. As a result, participants may have judged the CoT model more strictly, leading to a lower perceived naturalness despite its overall improvements in emotional expressiveness.

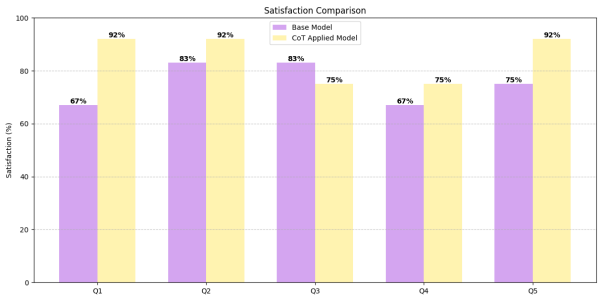

Figure 4: *Comparison of satisfaction ratings between the Base model and the CoT-applied model*

These findings suggest that while the CoT-applied model excels in interpreting emotional cues and delivering contextually rich responses, it also produces more complex speech patterns that may reveal minor imperfections. Nevertheless, the overall higher ratings in other categories indicate that the CoT model substantially enhances user experience.

5. Discussion

This study explored the potential of integrating AI-driven speech synthesis with Live2D animations to enhance emotional engagement in voice-based AI interactions. While previous research primarily focused on improving the naturalness and quality of speech synthesis [2], our study demonstrated that incorporating visual feedback can further deepen user immersion. Our prototype system, designed for Korean-language interactions, suggests that multimodal AI interfaces have the potential to enhance not only the perception of AI-generated speech but also the overall emotional connection between users and AI.

To gain preliminary insights into user experience, we conducted qualitative evaluations. People mentioned that AI models incorporating instructions led to more engaging and emotionally expressive interactions. However, key areas for further refinement were highlighted:

- Expanding the range of character expressions for more nuanced emotional feedback
- Improving the naturalness of synthesized Korean speech

These findings remain as future research topics for improving AI-driven voice interaction systems with greater immersion. While our study demonstrates the feasibility of integrating speech synthesis with animated visual elements, achieving a seamless and emotionally resonant experience requires additional effort. In particular, developing more personalized models by utilizing the functionality of Live2D Cubism to allow for fine-tuned adjustments of individual parts would be beneficial. Additionally, balancing response speed and the naturalness of synthesized speech remains a crucial challenge. A practical approach would be to initially leverage commercial APIs for rapid prototyping and then gradually refine the model through iterative training, optimizing it for specific emotional expressions in Korean speech synthesis.

Despite these challenges, our research provides foundational insights into the development of multimodal AI interactions. By combining high-quality speech synthesis with dynamic visual representations, we contribute to enhancing AI-driven emotional engagement. Future research should focus on advancements in speech emotion analysis and real-time animation technology to further refine AI voice companions. As AI continues to evolve in recognizing and responding to human emotions, we anticipate that AI-driven voice interactions will become increasingly natural and emotionally enriching.

6. Conclusion

In this study, we developed a web-based demonstration that facilitates empathetic conversations in Korean between users and AI by integrating Live2D characters and AI-generated speech. By implementing voice interactions through the gpt-4o-audio-preview model, we enhanced user immersion and naturalness, moving beyond static illustrations. Additionally, our system incorporates a robust STT component to maintain conversational context, ensuring continuity even in linguistically challenging scenarios.

To evaluate the effectiveness of the core speech model, we conducted emotion recognition tests under two conditions: with and without instructions. The results showed improved accuracy when instructions were provided, highlighting the importance of context in AI-driven emotion recognition.

Our findings demonstrate that integrating natural speech—supported by effective STT processing—and animated visual elements significantly enhances engagement and fosters more emotionally resonant AI interactions.

However, user feedback identified key areas for improvement, including expanding the range of character expressions for more nuanced emotional feedback and improving the naturalness of synthesized Korean speech. Addressing these challenges will be essential for future developments.

This study provides foundational insights into the development of multimodal AI interactions, demonstrating the added value of combining high-quality speech synthesis (complemented by robust STT) with dynamic visual representations in fostering emotional engagement. Future advancements in speech emotion analysis, real-time animation technology, and system stability will be crucial in further refining AI voice companions. As AI continues to evolve in recognizing and responding to human emotions, we anticipate that AI-driven voice interactions will become increasingly natural and emotionally engaging.

7. References

[1] H. Fei, H. Zhang, B. Wang, L. Liao, Q. Liu, and E. Cambria, "Empathyear: An open-source avatar multimodal empathetic chatbot," *arXiv preprint arXiv:2406.15177*, Jun 2024. [Online]. Available: https://arxiv.org/abs/2406.15177

[2] R. Chaudhury, M. Godbole, A. Garg, and J. H. Seo, "Humane speech synthesis through zero-shot emotion and disfluency generation," *arXiv preprint arXiv:2404.01339*, Mar 2024. [Online]. Available: https://arxiv.org/abs/2404.01339

[3] S. Rao, W. Xu, M. Xu, J. Leandro, K. Lobb, G. DesGarennes, C. Brockett, and B. Dolan, "Collaborative quest completion with llm-driven non-player characters in minecraft," *arXiv preprint arXiv:2407.03460*, Jul 2024. [Online]. Available: https://arxiv.org/abs/2407.03460

[4] R. E. Guingrich and M. S. A. Graziano, "Chatbots as social companions: How people perceive consciousness, human

likeness, and social health benefits in machines," *arXiv preprint arXiv:2311.10599*, Nov 2023. [Online]. Available: https://arxiv.org/abs/2311.10599

[5] OpenAI, "Hello gpt-4o," *OpenAI Research*, May 2024. [Online]. Available: https://openai.com/index/hello-gpt-4o/

[6] T. Guardian, "'hatsune miku has a special part in my heart': the 16-year-old pop sensation who does not exist," https://www.theguardian.com/music/2024/nov/26/hatsune-miku-john-cain-arena-melbourne-show-feature?utm_source=chatgpt.com, Nov 2024.

[7] R.-C. Chang, C.-S. Hung, B.-Y. Chen, D. Jain, and A. Guo, "Soundshift: Exploring sound manipulations for accessible mixed-reality awareness," *arXiv preprint arXiv:2401.11095*, Jan 2024. [Online]. Available: https://arxiv.org/abs/2401.11095

[8] L. Inc., "Live2d cubism: The industry standard for 2d real-time expression," https://www.live2d.com/en/cubism/about/, 2023.

[9] OpenAI, "Gpt-4o audio preview: Multimodal capabilities in conversational ai," *OpenAI Research*, Oct 2024. [Online]. Available: https://platform.openai.com/docs/guides/audio

[10] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," 2023. [Online]. Available: https://arxiv.org/abs/2201.11903

[11] TechTarget, "Gpt-4o vs. gpt-4: How do they compare?" https://www.techtarget.com/searchenterpriseai/feature/GPT-4o-vs-GPT-4-How-do-they-compare, 2024.

[12] R. M. Warren, "Perceptual restoration of missing speech sounds," *Science*, vol. 167, no. 3917, pp. 392–393, 1970.

[13] L. Inc., "Live2d sample data: Kei," https://www.live2d.com/en/learn/sample/, 2023.

[14] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[15] AIHub, "감정 분류를 위한 대화 음성 데이터셋," https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&dataSetSn=263, 2023.