

# AI English Conversation Coaching Platform in the Metaverse: Focused on Korean Users

Jaesung Park  
dept. of Computer Engineering  
Kumoh National Institute of  
Technology  
Gumi, Korea  
pjs7030@naver.com

Seongeun Yang  
dept. of Computer Engineering  
Kumoh National Institute of  
Technology  
Gumi, Korea  
banilla03090@gmail.com

Chaewon Yoon  
dept. of Computer Engineering  
Kumoh National Institute of  
Technology  
Gumi, Korea  
yce6345@kumoh.ac.kr

Jongwook Si  
dept. of Computer AI Convergence  
Engineering  
Kumoh National Institute of  
Technology  
Gumi, Korea  
jwsi425@kumoh.ac.kr

Yuchul Jung  
dept. of Artificial Intelligence  
Engineering  
Kumoh National Institute of  
Technology  
Gumi, Korea  
jyc@kumoh.ac.kr

Sungyoung Kim  
(\*Corresponding Author)  
dept. of Computer Engineering  
Kumoh National Institute of  
Technology  
Gumi, Korea  
sykim@kumoh.ac.kr

**Abstract**— In Korea, English education primarily focuses on exam-oriented learning, making it challenging to improve practical conversational skills. Additionally, Konglish or Koreanized English significantly hampers improving English conversational abilities. To solve this problem, this paper proposes a metaverse platform that offers English conversation functionalities in various scenarios, specifically targeting Korean users. The virtual environment for English conversation is created using the Unity Engine, leveraging the high performance and graphic quality provided by the URP. Digital humans are utilized as NPCs to enhance realism. The proposed platform allows users to select from various locations (City, Subway, Cafe, School) and engage in English conversations with NPCs based on the chosen setting. This English conversation functionality leverages GPT-based AI to provide interview simulation. Upon completion of a conversation session, the AI generates a conversation report to provide feedback to the user. This report offers a detailed analysis of the user's weaknesses and areas for improvement, contributing to the enhancement of their English conversational proficiency. Consequently, the proposed platform delivers an immersive English learning environment, facilitating effective English education through realistic and engaging experiences.

**Keywords**—AI, Metaverse, English Conversation, GPT, Speech

## I. INTRODUCTION

English is one of the most widely used languages globally, serving as a primary medium for international business, education, scientific research, and cultural exchange. In this context, effective English communication skills are essential for individual career development and for gaining a competitive edge internationally.

In Korea, the importance of English is further emphasized. Expanding into global markets, increasing cooperation with foreign companies, and extending international exchanges make English proficiency a necessary qualification. However, English education in Korea primarily focuses on exam-oriented learning, often disconnected from improving practical communication skills. In reality, when faced with situations requiring English usage, many Koreans experience a lack of confidence and a deficiency in practical conversational abilities. In particular, Konglish or Koreanized

English influences Koreans' spoken English [1]. Research on enhancing English conversational proficiency via V-logs or YouTube is ongoing [2]-[3]. However, a notable drawback when leveraging these platforms for language acquisition is that they do not accurately mirror real-life dialogues. This issue primarily arises from the content of these videos, which is skewed towards unidirectional information delivery. Genuine interpersonal communication is inherently bidirectional and evolves contextually, posing challenges in fully grasping conversational language's intricacies and subtle nuances. To address these issues, there is a growing need for platforms in the metaverse environment that enable learning English conversation in realistic scenarios. Various research is currently being conducted through metaverse platforms, which offer the significant advantage of allowing users to experience high realism [4]-[5]. Metaverse platforms are increasingly being leveraged for educational purposes, providing immersive experiences that closely mimic real-world interactions. Specifically, in English education, these platforms enhance language skills and interaction through immersive learning opportunities. Eduson[6] utilizes Extended Reality (XR) to enrich the learning process with live lectures by native speakers. In a similar vein, Immerse[7] offers Artificial Intelligence(AI)-driven conversational practice and voice chats with native speakers through its VR-based platform. VirtualSpeech[8] combines Virtual Reality(VR) and online training to create realistic virtual environments, providing AI-powered feedback on speaking skills across various scenarios. These are various examples of metaverse technologies being applied and serviced in the field of education.

Platforms utilizing metaverse technology provide environments that enable user interactions similar to real life, offering opportunities to use English in various situations ranging from everyday conversations to business meetings and cultural contexts. These simulations allow users to experience different real-life scenarios, enhancing the practicality of language learning and boosting confidence. While realism is important in platforms, what's more crucial for enhancing English conversational skills is the technology that provides accurate judgments and feedback. To address this issue, research into improving conversational abilities

using AI has been ongoing [9]-[12]. Therefore, by integrating the metaverse with AI, there is potential to solve all these issues.

The proposed method leverages GPT4[13]-based AI to interactively communicate with users, providing real-time feedback to improve conversational skills in metaverse environments. Users engage in natural dialogues using AI technology, which offers linguistic corrections and feedback. The system provides language processing capabilities for coaching, providing appropriate responses and corrections tailored to different conversational contexts. The features of the proposed method are as follows:

- **Realistic coaching with digital humans:** AI-driven coaching instructors provide realism, supporting effective language learning through real-time interaction.
- **GPT4[13]-based English conversation coaching:** Responds in real-time according to dialogue scenarios, improving conversational skills. User's speech is converted to text for input of GPT4[13], and responses are converted back to speech for the user.
- **Virtual English conversation content for various scenarios:** Users can enhance their conversational skills through a wide range of scenarios that mimic real situations.
- **Analysis reports based on conversation analysis:** After each conversation, user's utterances are analyzed to identify specific areas for improvement.

Introducing a metaverse environment for English conversation tailored to Korean users provides a language learning environment similar to real life allowing learners a more realistic and immersive experience. This approach redefines the direction of English learning and contributes to overcoming the limitations of traditional learning methods. This platform is effective due to the use of AI for feedback, response, and a realistic metaverse environment. It provides an immersive learning environment that enhances user concentration and offers conversational content in various scenarios similar to real-life situations, allowing users to learn practical English expressions that can be used in everyday life. Although the service was initially developed for Korean users, this platform can also be extended to non-English-speaking users. This platform allows users from various countries to achieve the same learning benefits, creating a global English learning environment.

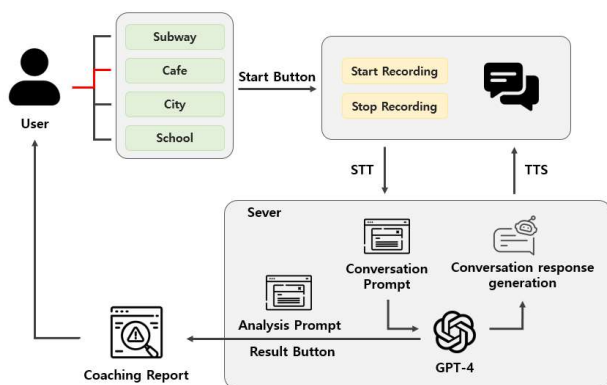


Fig. 1. Overall System of Proposed Method

This paper is structured as follows: Section II reviews AI prompt engineering and its application in English language education, Section III explores voice processing technologies, Section IV discusses the construction of metaverse environments for educational purposes, and Section V offers conclusions and recommendations for future research based on our findings. Ultimately, this paper proposes a metaverse-based platform tailored for enhancing English conversation skills.

## II. ENGLISH CONVERSATION COACHING TECHNOLOGY

The conversational coaching system proposed in this paper, as illustrated in Fig. 1, is designed to enhance English language proficiency by immersing users in various realistic conversational settings such as a Subway, Cafe, City, or School. Users initiate their experience by selecting a conversational objective from the interface, which then transports them to a virtual conversation space tailored to the selected scenario. Once the desired scenario is activated, users signal the start of the conversation by pressing the 'Start' button. This action initiates TCP socket communication between the user's device and the server, allowing for the exchange of data. Users interact with the system by recording their speech using the start and stop buttons. The audio is captured in WAV format, which preserves the quality of the speech for accurate processing. The server receives the WAV audio files and processes them through speech-to-text technology to transcribe the spoken words into text. Utilizing a conversational generation prompt, the server generates appropriate responses based on the dialogue context. These responses are converted back to audio in WAV format using Text-to-Speech and conveyed to the user, facilitating a seamless conversational flow. After at least three exchanges, a 'Results' button becomes available, giving users the option to review detailed analyses of their conversational interactions. Users can continue engaging in the conversation and click the results button as needed. Upon clicking the 'Results' button, the server analyzes the accumulated conversational data using a specific analysis prompt. It assesses various aspects of the user's performance, including fluency, accuracy, and appropriateness of responses. The comprehensive coaching results are then delivered to the user, providing valuable feedback aimed at enhancing conversational skills.

### A. Generation of conversation answers

Conversational practice can be conducted in four environments: Subway, City, Cafe, and School. Effective conversational responses are generated through Prompt Engineering using GPT-4 to facilitate natural dialogues with users. It is crucial for the AI to understand the context of the conversation through the user's questions and generate appropriate responses accordingly.

Prompts tailored to various conversational settings are provided to ensure that the AI produces relevant conversational responses. The prompts help the AI to comprehend the user-selected environment and generate situation-appropriate dialogue responses accurately. The AI model, acting as an English conversation tutor, creates suitable everyday conversational responses adjusted to the user's proficiency level to enhance their English speaking skills. The prompts for generating conversational responses include conditions to ensure the dialogue progresses naturally. All responses are generated in English, and conditions are set considering that some reactions might be less fluent

depending on the user's conversational ability. Additionally, since the conversation needs to be continuous, the generated responses are restricted to 1-2 sentences to keep them concise. To minimize hallucinations where the generative AI model produces irrelevant responses, cumulative conversation data with the user is provided, and the AI is restricted to generating only responses relevant to the user's questions. Fig. 2 shows an example of a conversational response generation prompt for the category of School.

A brief introduction to each environment and a predefined role for the AI is set to assist the AI in understanding each conversational environment and performing its role appropriately. In Subway, the AI is explicitly described as engaging in dialogues within a subway station or on a subway train, taking on roles such as a station worker or a passenger, and discussing events within the subway setting. In the Cafe, the AI assumes the role of a cafe staff member or patron, engaging in conversations typical of a cafe setting. In the City, the AI is designated to take on roles such as a police officer or a civilian, generating responses for incidents arising from requests for help on the streets. For the School setting, the conversations are confined to scenarios occurring within a university, where the AI can take on roles such as a student, friend, professor, or school staff, generating appropriate responses to the user's queries. Thus, within the roles defined

for each environment, the AI analyzes the conversational situation through the user's responses and generates suitable replies tailored to the specific role.

Fig. 3 illustrates an example of a conversation with a user in the Subway. As shown in the figure, when the AI receives information about the conversational setting from the user, the related Prompt is activated, providing the AI with basic descriptions of the conversational location and role information. In this case, the AI, taking on the role of a subway passenger, provides the user with information such as the arrival time at Gangnam station, the quickest exit, and details of a nearby cafe.

### B. Generation of Coaching Report

User conversational abilities are analyzed, and coaching reports are generated using prompt engineering techniques with GPT-4. Since the proposed coaching platform is targeted at services in Korea, unlike the conversational response generation process, which must be conducted in English, the conversational coaching reports are to be written in Korean (The prompt is translated into English to facilitate the reader's understanding). Therefore, conditions for generating reports in Korean have been added, and upon receiving a trigger word "evaluation" from the user, the generative AI model produces analysis results based on the user's conversational data.

```
Please generate your response in accordance with the following conditions:

###
<<Conversation Conditions>>
1. Since the user's English conversational skills are being assessed, all conversations should be conducted in English.
2. Given that each user's conversational skills may vary, be understanding of potential awkwardness in responses.
3. As the input text is generated through voice recognition, if there are typos, interpret and adjust the words according to the context as needed.
4. Keep responses brief, between 1-2 sentences.
5. Since the conversation takes place within a school setting, assume the role of a school student, friend, professor, or school staff as appropriate, and provide suitable responses to incidents occurring within the school.
6. Analyze the user's questions and assume an appropriate role in the conversation.
```

Fig. 2. Prompt of Conversation Response Generation for the School Category

```
>> User: Subway.
>> User: Excuse me, but is this the subway to Gangnam station?
>> Answer : Yes, this subway goes to Gangnam station. Make sure to get off at the 7th stop.
>> User: Oh great, I need to get off at Kangnam station. What station is it?
>> Answer : We are currently at Hongdae station. Gangnam station is 7 stops away from here.
>> User: Thanks! And do you know any good cafes close to Gangnam station?
>> Answer : Sure, there are many great cafes in Gangnam. One of them is "Cafe de Paris", which is known for its fruit-filled desserts and coffee.
>> User: Now, watch the closest exit to Gangnam station.
>> Answer : Sure, the closest exit to Gangnam station is Exit 1. It leads to the main road where you can find many shops and cafes.
>> User: This is really helpful. Thank you.
>> Answer : You're welcome! If you have any other questions, feel free to ask. Safe travels!
```

Fig. 3. Example of Conversation with User in the Subway

```
Please conduct the interview evaluation based on the following conditions:

###
<<Interview Response Evaluation Conditions>>
1. Since this is a Korean service, all responses, including analysis results, should be written in Korean.
2. If the response "평가" (evaluation) is requested, evaluate the user's deficiencies, areas for improvement, and suggestions for enhancement based on the English conversation data provided so far.
3. Correct any awkward contexts or words in the conversation, and point out and correct any incorrect tenses or grammatical errors.
4. If possible, suggest everyday conversational expressions that are commonly used or frequently spoken, and make the user's responses sound more natural.
```

Fig. 4. Prompt for Generating a Conversational Coaching Report

```
1. The user used an appropriate expression when making an order: "I'd like to order two coffees please." This phrase is very natural.

2. When requesting a specific type of coffee, the user said "Milk coffee please." While this is not grammatically incorrect, in English-speaking countries, the terms 'Cafe Latte' or 'Latte' are more commonly used. Thus, "I'd like two lattes, please." would be more natural.

3. The phrase "How much does it cost?" is not wrong, but in this context, "How much is it?" or "How much are they?" would be more suitable. 'Cost' is used when asking about the price of an item itself, whereas 'is' or 'are' are more common when inquiring about the price of items already ordered.

4. "Would I have this coupon credit please?" is slightly awkward. In English-speaking countries, "Could I use this coupon?" or "Can I apply this coupon?" are more natural expressions.

The user has basic conversational English skills and is particularly adept at using appropriate expressions when ordering. However, there is a need for more natural English expressions in some instances. Practicing commonly used English phrases in everyday situations would be helpful.
```

Fig. 5. Example of Conversation Coaching Reports for User in the Cafe

The prompt for generating a conversational coaching report is analogous to Fig. 4. To provide practical coaching analysis reports to users, the conversational coaching report generation prompt includes an analysis of user responses and essential elements of the report. The coaching report identifies any awkward usage of context or words during the conversation, provides feedback to correct them with the appropriate words, highlights these errors, and suggests corrections if incorrect tenses or grammar were used. Additionally, based on the conversational record, conditions are set to assess areas for improvement, weaknesses, and suggestions. The proposed conversational coaching platform focuses on enhancing conversational skills rather than just studying English, therefore it suggests modifications that encourage the use of expressions commonly used in everyday conversations, allowing users to develop more natural conversational abilities. Fig. 5 shows an example of creating a conversational coaching report in the Cafe. As seen in the figure, the coaching report provides analysis and feedback by correcting grammatical errors and awkward expressions in the user's conversational records and suggesting modifications to align with common expressions used in English-speaking locales.

### III. SPEECH RECOGNITION AND SYNTHESIS FOR INTERACTION WITH USER

Speech technology is crucial in maximizing the educational effectiveness of English conversational coaching systems. These systems allow users to converse naturally with NPCs, focusing on improving real-life communication skills. Users can initiate or continue conversations using speech, practicing pronunciation, intonation, and emotional expression in real time. This paper uses speech recognition and synthesis models to enable speech-based communication with the GPT-4 model. Speech technology allows users to communicate directly through their speeches rather than relying solely on text input. This approach enhances user experience, increases the accessibility of technology, and improves the naturalness of the interface. These benefits are critical factors in significantly enhancing the efficiency and satisfaction of users when utilizing English coaching. Additionally, using the GPT-4 model helps mitigate some misrecognition by the speech recognition model, demonstrating a complementary feature where each system supports and enhances the other.

#### A. Speech Recognition

Speech recognition is the technology that converts audio signals into text. It processes the user's speech through machine learning algorithms and deep neural networks. Most speech recognition models consist of acoustic and language models. The acoustic model analyzes the physical properties of sound to identify individual phonemes, while the language model understands the arrangement of words and grammatical structures for more accurate text conversion. These models are trained on large volumes of speech data, allowing them to accurately recognize different languages, dialects, speeds, and pronunciations. Speech recognition technology can interpret user commands in real-time, generate appropriate responses, or provide information. This is particularly crucial for interactive platforms like virtual assistants, customer service automation, and the coaching system discussed in this paper. Additionally, this technology enhances accessibility and improves the user experience. In this paper, the speech recognition functionality is implemented using Whisper[14].

#### B. Speech Synthesis

Speech synthesis models convert text to speech, adding specific speaker characteristics or emotions to the generated speech. During this process, text and speech are matched using a training dataset, and the model interprets the pronunciation of phonemes and the prosody and rhythm of sentences to train the encoder and decoder. In the text-to-speech conversion process, text first undergoes preprocessing to break it down into phoneme units, and vectors are then created to match each vocal feature. The decoder converts these vectors into a Mel spectrogram. Considering that human hearing is more sensitive to lower frequencies and less so to higher frequencies, the Mel spectrogram applies the Mel scale to reflect this. The Mel scale nonlinearly transforms frequencies to mimic human auditory characteristics. The transformed data produces a spectrogram via STFT(Short-Time Fourier Transform) and then applies the Mel scale to complete the Mel spectrogram. This process effectively captures various characteristics of real speech. Subsequently, a vocoder converts the Mel spectrogram back into speech. In this paper, the acoustic model and vocoder are trained together using VITS[15] to enhance speech quality.

### IV. ENGLISH CONVERSATION COACHING PLATFORM IN METAVERSE

This paper proposes a metaverse platform that provides English conversation functionalities in various scenarios, utilizing Unity Engine for creating virtual environments. While numerous game engines are available, Unity stands out due to its free usage, relative ease of learning and operation, and robust development capabilities, making content creation straightforward. Consequently, this study leverages Unity Engine to construct the virtual environment.

Unity's default rendering pipeline is the Built-In Render Pipeline. In contrast, the SRP (Scriptable Render Pipeline) allows for rendering control and customization via C# scripts. Unity offers two built-in SRPs: URP (Universal Render Pipeline) and HDRP (High Definition Render Pipeline). This paper employs URP due to its superior performance and enhanced graphical quality. URP, more flexible and scalable than the Built-In Render Pipeline, provides advanced graphics optimized for various platforms, achieving more realistic visuals. Thus, the URP pipeline is utilized to create realistic content.

The proposed technology aims to enhance realism in English conversations by employing digital humans and high-quality Unity assets. Digital humans are virtual beings created using digital technology to resemble human appearance and behavior closely. Various techniques, such as 3D modeling, animation, and voice synthesis, achieve realistic appearance and body movements. The digital humans are created using MAYA 2020 and incorporate Blend Shapes. Blend Shapes is a technique used to alter a character's facial expressions or mouth shapes, simulating changes in facial expressions and appearance. The digital humans in this study are finely adjustable with Blend Shapes applied from head to toe.

To enhance realism, lip-sync technology is applied, ensuring that the digital human's mouth movements match the spoken audio. Among various lip-sync technologies, uLipsync[16] is adopted to implement lip-sync for digital humans. uLipsync analyzes speech into vowel sounds A, I, U, E, O, and adjusts the Max Weight values of the Blend Shapes corresponding to these vowel sounds to replicate the

appropriate mouth shapes for the digital human. Max Weight represents the maximum intensity of the transformation when Blend Shapes are applied, ranging from 0 (no application) to 1 (maximum application). This ensures natural lip-syncing. Additionally, uLipsync supports real-time voice analysis and microphone input, facilitating smooth English conversations with the digital human. The presence of Blend Shape functionality in the digital human is crucial for achieving realistic lip-syncing.

In this paper, digital humans are utilized as NPCs to conduct English conversations. The scenario-based scene composition is illustrated in Fig. 6. The user begins in the Start Menu scene, where they can select one of several conversation locations. After selecting one of the four locations, the user engages in an English conversation with the NPC placed in the chosen scene. Upon concluding the conversation, the Result scene analyzes the user's responses based on the conversation with the NPC.

The proposed method allows for the selection of English conversation content suitable for various locations such as city, subway, cafe, and school, as depicted in Fig. 7. The city scene is implemented using the 'Easy Modular City'[17] asset, while the subway scene is based on the 'Sample Subway Scene' from 'City Subway-Collection'[18]. The cafe scene offers a cozy interior created with the 'Coffee Shop Interior' asset[19], and the school scene captures the campus atmosphere using the 'School Scene' asset[20].

In Fig. 8, the user views various locations from a first-person perspective. The user can navigate the scene using the keyboard's WASD keys and smoothly rotate the view by moving the mouse, enhancing natural conversation scenarios and realism. Approaching an NPC within 2.0 units and pressing the spacebar initiates an English conversation with the NPC. The digital human NPCs are situated in each location, and a virtual environment appropriate to the selected area has been created. The NPCs, resembling real humans, and the virtual environments matching the chosen locations provide users with realistic immersion.

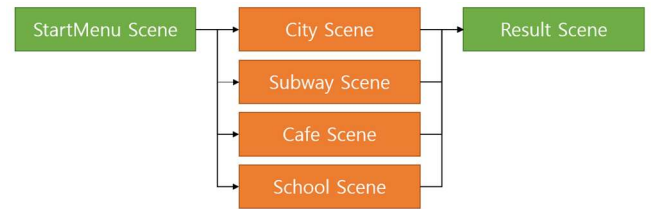


Fig. 6. Scene Composition Based on Contents Scenario

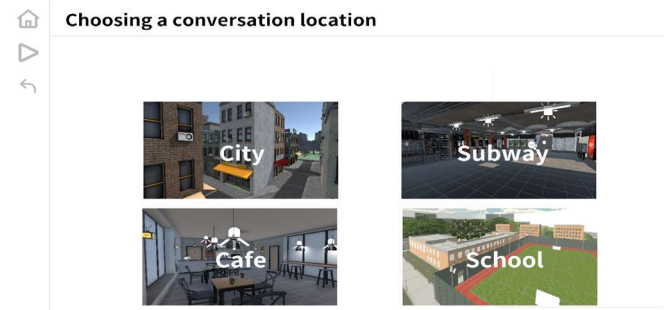


Fig. 7. Interface of StartMenu Scene

Pressing the spacebar reveals three buttons. Pressing the first button, the Start Button, records the user's speech saying "City", "Subway", "Cafe", or "School". It sends the recording to the server to prepare for conversations appropriate to each location. After pressing the Start Button, the user presses the second Start Recording button to record their speech, saving it as a WAV file. Upon completing their response, the user presses the third Stop Recording button to send the saved WAV file to the server. The server then responds with a WAV file containing the answer to the user's question, which is automatically assigned in real-time to the Audio Clip section of the NPC's Audio Source, playing the sound. Audio Source plays audio clips in the scene, while Audio Clip contains the audio data used by the audio source. The user continues the conversation by pressing Start Recording again and repeats the process by pressing Stop Recording.

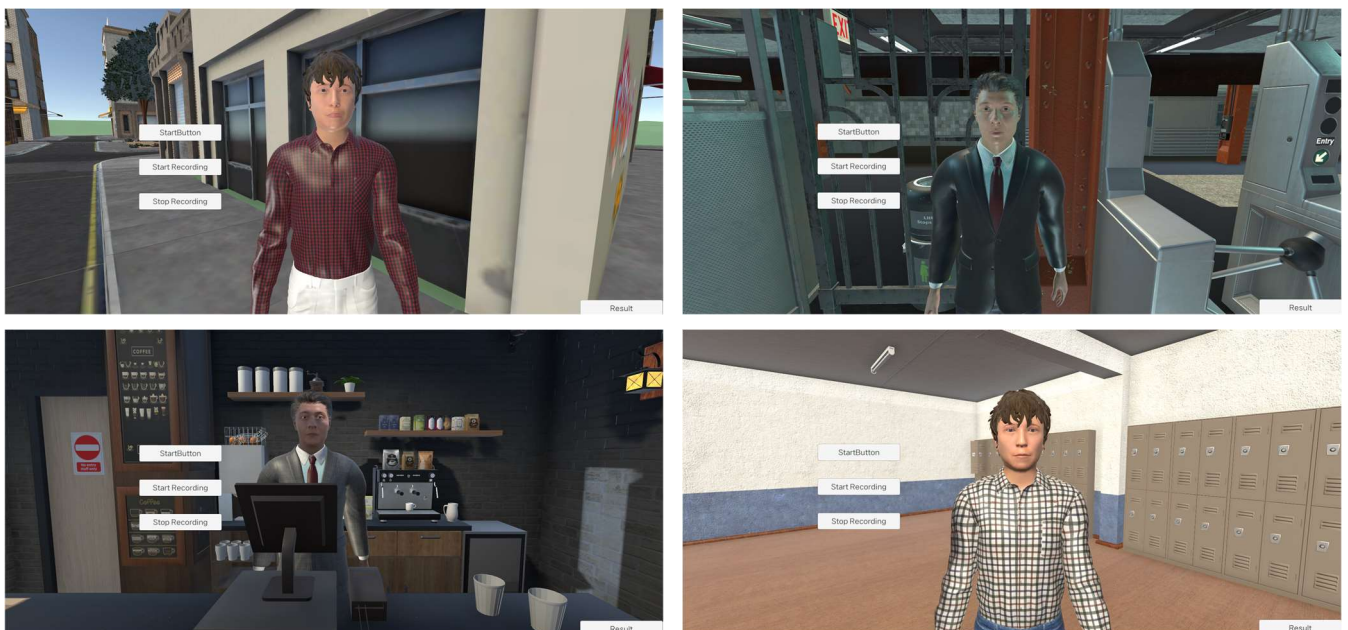


Fig. 8. Scene Interface of the English Conversation Contents with NPC (Top to Bottom: City, Subway, Cafe, School)

## Conversation Coaching Report Result

1. The user used an appropriate expression when making an order: "I'd like to order two coffees please." This phrase is very natural.
  2. When requesting a specific type of coffee, the user said "Milk coffee please." While this is not grammatically incorrect, in English-speaking countries, the terms 'Cafe Latte' or 'Latte' are more commonly used. Thus, "I'd like two lattes, please." would be more natural.
  3. The phrase "How much does it cost?" is not wrong, but in this context, "How much is it?" or "How much are they?" would be more suitable. 'Cost' is used when asking about the price of an item itself, whereas 'is' or 'are' are more common when inquiring about the price of items already ordered.
  4. "Would I have this coupon credit please?" is slightly awkward. In English-speaking countries, "Could I use this coupon?" or "Can I apply this coupon?" are more natural expressions.
- The user has basic conversational English skills and is particularly adept at using appropriate expressions when ordering. However, there is a need for more natural English expressions in some instances. Practicing commonly used English phrases in everyday situations would be helpful.



Fig. 9. Example of Conversation Coaching Report

When the user wishes to end the English conversation, they press the result button at the bottom right of Fig. 8 to move to the resulting scene. In the resulting scene, pressing the result button next to the conversation coaching report initiates the analysis of the conversation with the NPC, providing an assessment of the user's overall conversation skills and areas for improvement. The report's content shown in Fig. 9, is displayed on the screen by receiving the inferred results from Fig. 5 through the server, allowing the user to view the results directly.

### V. CONCLUSIONS

This paper introduces a platform leveraging metaverse technology to simulate realistic and diverse English conversation scenarios for language proficiency enhancement. It allows learners to augment their linguistic capabilities through virtual experiences, from casual dialogues to professional interactions. Additionally, real-time feedback in the virtual environment is poised to enhance users' proficiency in practical language applications significantly. This metaverse-based English coaching system pioneers a new paradigm in language education by offering a flexible learning environment.

Future research should delve deeper into the efficacy of this platform and expand the system to include tailored coaching techniques for various languages and scenarios. This would aim to effectively improve users' conversational skills across multiple languages. Such studies are expected to make significant contributions to global language education.

### ACKNOWLEDGMENT

This work was supported by Metaverse Lab Program funded by The Ministry of Science and ICT (MSIT), and The Korea Radio Promotion Association (RAPA)

### REFERENCES

- [1] Y. Fayzakhmanova, "Koreanized English words from perspectives of Korean-English language contact," *Asian Englishes*, Vol. 18, No. 3, pp. 216-231, 2016.
- [2] E. Meinawati, D. D. Harmoko, and N. A. Rahmah, "Increasing English speaking skills using YouTube," *Polyglot: Jurnal Ilmiah*, Vol. 16, No. 1, pp. 1-13, 2020.
- [3] B. Mandasari and D. Aminatun, "VLOG: A Tool To Improve Students English Speaking Ability At University Level," *International Conference on Linguistics, Language Teaching, Literature and Culture*, Jul 2019.
- [4] J. Si, S. Yang, D. Kim, and S. Kim, "Metaverse Interview Room Creation with Virtual Interviewer Generation Using Diffusion Model," *IEEE Asia-Pacific Conference on Computer Science and Data Engineering*, Dec 2023.
- [5] C. Yoon, S. Yang, J. Park, J. Si, Y. Jung, and S. Kim, "Metaverse Virtual Interview Platform Leveraging Generative AI and Speech Recognition," *Journal of Korean Institute of Information Technology*, Vol. 22, No. 6, pp.163-173, Jun 2024.
- [6] Eduson English, <http://edusonenglish.com/>
- [7] Immerse, <https://www.immerse.com/>
- [8] VirtualSpeech, <https://virtualspeech.com/courses/english-for-business>
- [9] T. Akiyama, K. Arai, T. Kobayashi, and N. Nakamura, "Research on AI for English Conversation for Elementary School Students," *IEEE 10th Global Conference on Consumer Electronics* pp. 503-504, Oct 2021.
- [10] H. Yang, H. Kim, J. Lee, and D. Shin, "Implementation of an AI chatbot as an English conversation partner in EFL speaking classes," *ReCALL*, Vol. 34, No. 3, pp. 327-343, 2022.
- [11] S. Ruan, L. Jiang, Q. Xu, Z. Liu, G. M. Davis, E. Brunskill, and J. A. Landay, "Englishbot: An AI-Powered Conversational System for Second Language Learning," *International Conference on Intelligent User Interfaces*, pp. 434-444, Apr 2021.
- [12] I. P. Hapsari and T. T. Wu, "AI chatbots learning model in English speaking skill: Alleviating speaking anxiety, boosting enjoyment, and fostering critical thinking," *In International Conference on Innovative Technologies and Learning*, pp. 444-453, Aug 2022.
- [13] J. Achiam, B. McGrew *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*.
- [14] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *In International Conference on Machine Learning*, pp. 28492-28518, Jul 2023.
- [15] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," *In International Conference on Machine Learning*, pp. 5530-5540, Jul 2021.
- [16] uLipsync, <https://github.com/hecomi/uLipSync>
- [17] City Subway-Collection, <https://assetstore.unity.com/packages/3d/environments/city-subway-collection-168439>
- [18] Easy Modular City, <https://assetstore.unity.com/packages/3d/environments/urban/easy-modular-city-184669>
- [19] Coffee Shop Interior, <https://assetstore.unity.com/packages/3d/props/interior/coffee-shop-interior-209622>
- [20] School Scene, <https://assetstore.unity.com/packages/3d/environments/urban/school-scene-66006>