

Learning English conversation in a metaverse: The possibilities and challenges of world heritage tours

Eisaku Takahashi

Department of Regional Development, Takasaki City University of Economics, , eisaku@tcue.ac.jp

How to cite: Takahashi, E. (2025). Learning English conversation in a metaverse: The possibilities and challenges of world heritage tours. In Y. Choubasaz, P. Díez-Arcón, A. Gimeno-Sanz, V. Morgana, A. C. Murphy & F. L. Seracini (Eds.), *Advancing CALL: New research agendas - EUROCALL 2025 Short Papers*. <https://doi.org/10.4995/EuroCALL2025.2025.21245>

Abstract

This study explored the effects of voice quality and chatbot-based pre-task practice on learners' foreign language anxiety (FLA) and speaking behavior in a metaverse-based English conversation environment. Thirteen Japanese university students participated in a ten-session virtual tour program featuring guided conversations conducted entirely in English. Quantitative data were collected via pre- and post-session questionnaires, while qualitative data were obtained through open-ended responses and researcher observations. Findings revealed that, although many participants initially felt anxious and lacked confidence in speaking English, repeated engagement in the metaverse reduced resistance to speaking for several learners. In particular, the clarity and tone of the instructors' voices significantly influenced learners' sense of psychological comfort and willingness to participate. While chatbot practice received generally positive feedback, some learners reported usability issues, and anxiety about spontaneous speech remained. Additionally, learners showed limited interest in avatar customization, and no evidence of the Proteus effect was observed. These results highlight the importance of voice design in virtual language learning and suggest that immersive environments can support anxiety reduction through repeated, low-stress interactions. Further research is needed to examine how vocal characteristics and avatar identity influence learner psychology in metaverse-based English education.

Keywords: *metaverse-based English education; foreign language anxiety; chatbot; voice quality.*

1. Introduction

Foreign Language Anxiety (FLA), which refers to the anxiety learners experience when engaging in conversation in a foreign language, is widely recognized as a major obstacle in the language acquisition process (Krashen, 1982; Hojo, 1996). Due to such anxiety, many learners tend to avoid opportunities to speak English in front of others. Recently, English conversation learning in metaverse environments has attracted increasing attention as a potential solution to this issue (Li & Yu, 2023; Kaplan-Rakowski & Gruber, 2023; Takahashi, 2024). The metaverse offers a novel learning environment where learners can engage in experiences that closely simulate real-life situations within a virtual space. Through activities such as virtual tourism and dialogue via avatars, learners are able to improve their English skills while reducing the psychological burden typically associated with real-world interactions. This mode of learning has gained particular significance in the aftermath of the pandemic, during which opportunities for overseas travel and study abroad have been substantially limited. The metaverse holds considerable social value by enabling cultural experiences, such as visits to World Heritage sites, beyond physical constraints.

Takahashi (2024) notes that while the metaverse may help mitigate anxiety associated with face-to-face interaction, it also introduces new challenges, particularly the absence of nonverbal cues during communication. Mehrabian (1971) argued that 93% of human communication relies on nonverbal signals; elements such as facial expressions, eye contact, and gestures play a crucial role in understanding conversational intent. In the metaverse, however, the expressive capabilities of avatars are limited, which can hinder the identification of speakers and the interpretation of questions. Moreover, the “foreign language effect,” whereby thinking in a non-native language leads to reduced cognitive processing (Takano & Noda, 1995), may further inhibit learners’ ability to respond promptly. In such environments, where verbal information dominates the majority of communicative content, the quality of voice—particularly that of the speaker—can directly influence learners’ comprehension and sense of psychological comfort.

This study focuses on how the voice quality of instructors affects learners’ anxiety levels and fluency during conversational activities. As a preparatory step, participants engaged in chatbot-based practice sessions prior to the main conversation tasks, aiming to reduce psychological resistance to voice-based interaction. The insights gained from these practices are expected to inform the design of chatbot and narration voice features in future metaverse-based English learning environments, particularly in terms of audio settings.

The purpose of this study is to examine the effects of voice quality and pre-task practice on the reduction of foreign language anxiety and the enhancement of communicative competence in English within a metaverse learning environment. Specifically, the study investigates how learners’ impressions of the speaker’s voice contribute to their sense of psychological security and willingness to speak in a virtual space where non-verbal cues are limited. The research involved learners participating in a “World Heritage Metaverse Tour,” in which an English instructor acted as a tour guide. Prior to the tour, learners engaged in conversational practice with a chatbot to reduce their resistance to voice interaction and to facilitate practical preparation. The findings of this study are expected to serve as a foundational reference for designing effective chatbot voices and optimizing pedagogical dialogue in metaverse-based English education.

2. Materials and methods

2.1. Participants

Thirteen undergraduate students (3 male and 10 female), ranging from second- to fourth-year university students, participated in this study. Participation was voluntary, and all participants were fully informed of the purpose and procedures of the study prior to recruitment. At the time of participation, their English proficiency, as measured by the Minimal English Test (MET) and converted into TOEIC-equivalent scores, averaged 444.5, indicating that most participants had a lower-intermediate level of proficiency. To ensure ethical research practices, all personal information was anonymized, and participants were explicitly informed that the collected data would not be reused for secondary purposes. The study was reviewed and approved by the institutional ethics committee of the affiliated university and conducted in accordance with established research ethics guidelines.

2.2. Learning environment

The English conversation activities in this study took place within a custom-designed metaverse environment. This environment was developed by SystemCreators Corporation, a technology company located in Takasaki City, Gunma Prefecture, Japan. This virtual space featured digital reconstructions of eight UNESCO World Heritage sites, and learners participated in guided virtual tours set in each location. Participants accessed the system from their personal computers to engage in the conversation activities. VR headsets were not used due to concerns about potential health issues such as dizziness and because PC-based access posed a lower barrier to participation. All communication during the activities was conducted entirely in English, with instructors guiding the tours. Participants joined the sessions either from their homes or from classrooms on the university campus.

2.3. Structure of the learning program

The learning program, titled the “World Heritage Metaverse English Conversation Tour,” was conducted over the course of ten sessions within the virtual heritage space. Each session lasted 60 minutes and consisted of an English-language guided tour followed by related conversation practice. The destinations visited in each session and the instructors in charge were as follows:

- | | |
|------------------------------------|------------------------------------|
| 1. Statue of Liberty (N) | 2. Uluru and Kata Tjuta (N) |
| 3. Taj Mahal (E) | 4. The Acropolis and Parthenon (N) |
| 5. The Great Pyramid of Giza (A) | 6. Machu Picchu (A) |
| 7. Vatican City (N) | 8. Sannai Maruyama Ruins (E) |
| 9. Statue of Liberty (revisit) (N) | 10. Taj Mahal (revisit) (N) |

* The letters in parentheses next to each session indicate the initials of the instructor responsible for that tour.

The instructional team consisted of two male instructors and one female instructor, who acted as both facilitators of dialogue and tour guides. All instructors are presently affiliated with M • Y Associates Ltd., in Takasaki City, Gunma Prefecture, Japan. One instructor accessed the metaverse space from South Africa, while the other two joined from Takasaki City in Gunma Prefecture, Japan. Each instructor provided commentary modeled on that of a local tour guide and actively encouraged dialogue with the learners.

Prior to each session, students completed pre-task practice using a chatbot. This chatbot engaged learners in question–response exchanges that included vocabulary and expressions likely to be used during the tours, allowing them to practice speaking through voice input. This pre-task activity was designed to reduce psychological resistance to spoken interaction within the virtual environment.

The program followed a consistent structure: pre-task practice → main session → instructor feedback. After each session, the instructor provided feedback to the researcher (the author), reporting on participants’ reactions, speaking patterns, and noteworthy instances of language use observed during the activity. These feedback reports were used as supplementary material for subsequent data analysis and interpretation.

2.4. Data collection methods

To gain a comprehensive understanding of English conversation activities in the metaverse environment, this study employed a mixed-methods approach combining both quantitative and qualitative data collection techniques. First, all sessions were recorded using PCs, and the recordings were stored as video data suitable for analyzing conversational flow, speech quantity, turn-taking patterns, and the content of instructor utterances. The recordings adopted a bird’s-eye perspective that captured an overview of the virtual tours, including spatial layouts, instructor movements, and the positioning of learner avatars. Given the structure of the program, interactions among participants were not expected; instead, the dialogue occurred in either one-on-one or one-to-many formats between instructors and learners.

In addition, participants were asked to complete questionnaires both before and after each session. The questionnaire, developed with reference to Kaplan-Rakowski and Gruber (2023), included items measuring foreign language anxiety, self-assessed English proficiency, impressions of the instructors’ voices, and comfort levels with communication in the metaverse environment. All items were rated on a five-point Likert scale (1 = strongly disagree, 5 = strongly agree), and the responses were quantified to calculate average scores for quantitative analysis. Open-ended response sections were also provided to collect qualitative data on psychological changes and subjective evaluations of audio-related factors throughout the program.

By integrating video recordings, questionnaire data, and observational notes, the study established a robust framework for examining linguistic behavior and psychological factors from multiple perspectives in the context

of English conversation activities conducted within a metaverse environment.

2.5. Data analysis methods

The collected data were analyzed using both quantitative and qualitative approaches. For the quantitative component, descriptive statistics (mean, standard deviation, and score differences) were calculated to examine changes in self-assessed communicative competence in English and levels of foreign language anxiety based on pre- and post-session questionnaires. Learners' impressions of the instructors' voices were analyzed using both the rating items and open-ended responses from the post-session questionnaire. Particular attention was paid to how vocal qualities, tone, and speaking style may have influenced learners' psychological responses. It should be noted that the video recordings were not used for analytical purposes in this study. For the qualitative component, content analysis was conducted on the open-ended responses from the post-session questionnaires, focusing on descriptions related to learners' speaking experiences. Particular emphasis was placed on comments addressing feelings of comfort or reductions in anxiety experienced during the activities. Observation notes taken by the researcher were also reviewed to identify patterns in avatar selection among participants, with the goal of studying how learners employed strategies of self-representation in an audio-centered virtual environment. By integrating these quantitative and qualitative data sources, the study sought to establish a basis for discussing how English conversation activities in a metaverse environment may influence learners' anxiety levels and motivation for language learning.

3. Results

3.1. Questionnaire results

Results from the pre-session questionnaire indicated that many participants experienced a noticeable degree of anxiety when speaking English. In particular, high average scores were observed for items such as "I have no confidence at all when speaking English" ($M = 3.46$), "I feel more nervous and anxious when I have to speak English than when I speak Japanese" ($M = 3.38$), and "I feel very afraid when I have to speak English without preparation" ($M = 3.08$). These results suggest that participants felt considerable anxiety in situations requiring spontaneous speech.

In contrast, items related to positive self-assessment showed relatively lower scores. Statements such as "I feel confident when speaking English" and "When I know in advance that I have to speak English, I feel at ease and relaxed" received lower ratings overall, indicating a generally low sense of self-efficacy regarding English speaking. The item "I don't mind making mistakes when I speak English" had an average score of 3.0, which suggests variability in anxiety levels across participants.

The post-session questionnaire also revealed continued anxiety about spontaneous speech. The item "I feel very anxious when I have to speak English without preparation" remained the highest ($M = 3.38$), and "When I had to talk about topics I had not prepared for in the VR environment, I felt nervous" also scored relatively high ($M = 3.15$). These results suggest that the chatbot-based pre-task practice may not have been sufficient to fully alleviate anxiety related to unprepared conversation.

However, a slight decrease was observed in the score for "I have no confidence at all when speaking English," which dropped from 3.46 (pre) to 3.15 (post). This change indicates that some learners may have gained a degree of confidence in speaking. These findings suggest that continued exposure to English in the metaverse environment may have contributed to reducing psychological barriers for at least some participants.

3.2. Perceived vocal qualities and ease of listening

In the post-session questionnaire, participants were asked to evaluate the clarity and ease of listening to the instructors' voices, who acted as tour guides during the activities. According to the aggregated results, the instructor labeled "N" was rated as having the easiest voice to understand by 11 participants, significantly

outperforming the other instructors. The next highest rating was given to instructor “A” (2 participants), while no participants selected the remaining instructor. These results indicate that learners perceived instructor N’s speaking style and vocal qualities as the clearest and most reassuring. “N” and “E” are male instructors from South Africa, and “A” is a female instructor from the United States.

Regarding preferred voice types, five participants selected “a low-pitched male voice” and five selected “a high-pitched female voice” as the most easily understandable. This was followed by “a high-pitched male voice,” chosen by two participants. These tendencies indicate that learners found either lower-frequency male voices or higher-frequency female voices to be easier to process. In virtual, non-face-to-face, and largely non-interactive environments, voice is the primary source of real-time information, and thus the frequency range and clarity of speech may have a significant impact on learner comprehension and perceived comfort.

It should be noted that no acoustic analysis was conducted on the recorded video data. All findings related to vocal qualities were derived exclusively from participants’ subjective evaluations as reported in the post-session questionnaire. While these findings reflect individual perceptions, they point to the potential pedagogical importance of selecting appropriate vocal qualities when designing audio-based instructional environments.

3.3. Evaluation of chatbot use

In this study, participants engaged in voice-based practice sessions with a chatbot prior to each metaverse activity. The effectiveness of these sessions was assessed through qualitative analysis of open-ended responses from the post-session questionnaire. These responses were categorized into four types: positive, negative, neutral, and mixed. The following trends were observed.

Positive evaluations were the most frequently reported (5 responses) and included comments such as “It was fun,” “It was useful,” and “I was glad to be able to practice in advance.” These responses suggest that the chatbot was perceived as a friendly and effective tool for preparing for spoken interaction.

Negative evaluations (2 responses) included remarks such as “Too much text made it difficult to read,” “The response wasn’t very good,” and “I didn’t understand how to use it.” These comments indicate that, for some participants, interface visibility and speech response accuracy may have posed usability challenges or sources of stress.

Although positive impressions outweighed the negative, the findings also highlight potential limitations in terms of visual clarity and user operability for certain learners. Moreover, the relatively high average score ($M = 3.15$) for the item “I felt nervous when I had to talk about unprepared topics in the VR environment” in the quantitative data suggests that the chatbot practice may not have fully alleviated anxiety related to spontaneous speech.

Taken together, these results indicate that improvements in the design of chatbot conversations—as well as enhancements in voice output and display layout—may contribute to more effective anxiety-reduction strategies in virtual language learning environments.

3.4. Behavioral trends based on observational notes

Observational notes recorded by the researcher during each session revealed several notable patterns regarding participant behavior and adaptation to the learning environment.

First, regarding avatar selection, multiple instances occurred in which all participants chose the same avatar, despite being given the option to freely select from six different avatars. This trend suggests that learners showed relatively little interest in the visual or role-based aspects of avatar customization and were more focused on voice-based communication rather than external self-representation. This study did not observe the “Proteus effect”—a phenomenon in which one’s virtual appearance influences behavior, as discussed by Yee, Bailenson, and Ducheneaut (2009).

Second, a clear trend emerged regarding the relationship between the number of sessions attended and the

frequency of speech. Observational data indicated that individual speaking frequency increased as participants attended more sessions. This suggests that increased familiarity with the virtual learning environment may have contributed to greater willingness to speak and enhanced learner confidence.

In addition, feedback provided by instructors to the researcher included comments such as “Participants spoke very little in the first session, but their responses increased over time” and “They began responding to questions more quickly.” These observations support the interpretation that repeated experience in a voice-based virtual environment facilitated behavioral adaptation and promoted learner participation in spoken interaction.

4. Discussion

This study investigated, from multiple perspectives, the impact of voice-based English conversation activities conducted in a metaverse environment on learners’ anxiety and speaking behavior. The findings derived from both quantitative and qualitative data suggest that, while many participants initially experienced anxiety about speaking English, their confidence and willingness to speak gradually increased through repeated participation in virtual tour sessions.

The majority of participants rated instructor N as having the “easiest voice to understand,” a particularly notable result. In fact, 11 of the 13 participants chose N, demonstrating their perception of his vocal tone and delivery as particularly clear and reassuring. Comments in the post-session open-ended responses included remarks such as “His voice was easy to follow and enjoyable,” suggesting that vocal clarity significantly reduced learners’ psychological burden. As Takahashi (2024) has pointed out, the absence of nonverbal cues in metaverse communication can contribute to misunderstanding and anxiety. In such an environment, where voice serves as the only real-time affective channel, auditory clarity may directly affect the success of interaction. Furthermore, the impression of the speaker’s voice is not merely a channel for transmitting information; it also plays a vital role in fostering comfort and willingness to participate. This aligns with Gobl and Ní Chasaide’s (2003) theory that vocal quality conveys emotional attributes. In a virtual environment with limited visual and physical feedback, voice becomes a core element in constructing social relationships.

Another key finding was that learners’ frequency of speaking increased with the number of sessions attended. This trend, identified through observational notes, revealed that participants became more proactive in taking speaking turns over time. Early sessions were characterized by silences or delayed responses, but later sessions displayed increased spontaneous participation. These patterns suggest that learners became psychologically and operationally more accustomed to the virtual environment. As Takahashi (2024) has proposed, metaverse environments may serve to alleviate FLA. The observation that repeated participation reduced resistance to speaking supports this hypothesis. In particular, the lack of visible social judgment—such as gaze or peer evaluation commonly felt in face-to-face environments—may have contributed to behavioral shifts in this voice-centric virtual space.

The Proteus effect, which suggests that one’s virtual self-image influences behavior (Yee, Bailenson, and Ducheneaut, 2009), did not manifest itself in this study, as participants showed limited interest in avatar selection. Despite being free to choose from six avatars, participants often selected the same one. This suggests that learners prioritized audio communication over visual self-expression, possibly due to the voice-only nature of the

interaction. It was oribably the one that was the easiest choose (Figure 1).



Figure 1. Select an avatar

The pre-task chatbot practice was generally perceived as beneficial. Positive feedback included remarks such as “It was fun” and “It helped,” indicating that the chatbot played a useful role in preparing learners for speaking activities. However, some participants reported issues such as “too much text,” “poor responsiveness,” and “unclear interface,” suggesting usability challenges. While the chatbot fulfilled its role as a basic tool for voice-based practice, it did not eliminate anxiety about spontaneous speech. Improvements in functionality remain necessary.

The findings of this study offer several important implications for the design of VR-based English education. In particular, the impact of instructors’ vocal characteristics on learner engagement in voice-centric environments highlights the importance of designing speech output that promotes emotional comfort. This insight is relevant for the development of AI-driven chatbots and voice-based instructional materials. Furthermore, the observed reduction in speaking anxiety through repeated participation suggests that virtual environments may serve as effective platforms for supporting learners who experience FLA.

Additionally, the immersive virtual environment—designed to replicate real-world experiences through interactive tours—served as a motivational factor for learners. The gamified, tourism-based structure helped engage learners not only cognitively but also emotionally, providing opportunities for active language use. In this respect, the study has broader applications in the field of educational technology. Moving forward, the design of VR English learning environments should integrate sensory elements such as visual presentation and vocal quality, along with scaffolding systems tailored to learners’ psychological states.

That said, several limitations must be acknowledged. First, most participants had relatively low English proficiency, with TOEIC-equivalent scores below the intermediate level. Therefore, the findings may have limited generalizability to more advanced learners. Second, the sample size was small ($N = 13$), which poses limitations in terms of statistical reliability and applicability to diverse learning styles. Third, the metaverse environment used in this study was accessed via PC screens, not immersive VR headsets. Thus, the influence of bodily immersion and spatial realism on speaking anxiety and learner motivation warrants further investigation.

Future studies should focus on the design and evaluation of chatbot voice features. For example, beyond pitch and speaking rate, it is worth examining how speaker identity—such as whether the voice belongs to a native English speaker or a Japanese speaker—affects learners’ psychological comfort and willingness to speak. Re-examining the Proteus effect is also a promising line of inquiry. For instance, would learners gain more confidence in speaking if their avatars resembled native English speakers? Would a chatbot with a Japanese appearance and voice lower

the psychological threshold for conversation? Addressing these questions may be essential for the future development and implementation of virtual language learning systems.

In this regard, the present study serves as a starting point for further empirical research and applied development in the field of immersive language education.

5. Conclusion

This study aimed to investigate how voice quality and pre-task practice using a chatbot influence learners' foreign language anxiety and speaking behavior in English conversation activities conducted within a metaverse environment. The findings confirmed that participants experienced anxiety and a lack of confidence when speaking English; however, repeated engagement in the virtual sessions appeared to alleviate some of this resistance to speaking over time.

In particular, the voice quality of instructors was found to play a significant role in fostering psychological comfort. Voices rated as "easy to understand" were associated with increased learner willingness to speak, highlighting the importance of vocal characteristics in voice-centric virtual environments. Additionally, the observed increase in speaking frequency with more session participation suggests that familiarity with the environment may help reduce psychological barriers. In contrast, learners showed little interest in avatar customization, and no evidence of the Proteus effect—whereby visual self-representation influences behavior—was observed. While chatbot-based pre-task practice received generally positive feedback, it was not sufficient to fully eliminate anxiety about spontaneous speaking, indicating the need for further improvements in functionality.

These findings provide valuable advice regarding the design of metaverse-based English learning environments that prioritize vocal design and suggest directions for educational interventions aimed at reducing foreign language anxiety. However, several limitations should be noted: the small sample size and the predominance of lower-proficiency learners limit the generalizability of the findings, and the study environment, while virtual, did not employ immersive VR headsets, which may have influenced learners' sense of presence and engagement.

Future research should explore how specific vocal qualities and speaker attributes (e.g., Japanese-accented vs. native English-accented voices) influence learners' psychological responses. Additionally, the relationship between avatar appearance and speaking confidence—especially in relation to the Proteus effect—warrants further investigation. Addressing these questions will contribute to the development of more effective, psychologically attuned metaverse-based English learning environments.

Acknowledgments

This work was supported by Takasaki City University of Economics Research Encouragement Grant 2024 No. 7.

References

- Gobl, C., & Ní Chasaide, A. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*, 40(1–2), 189–212. [https://doi.org/10.1016/s0167-6393\(02\)00082-1](https://doi.org/10.1016/s0167-6393(02)00082-1)
- Hojo, R. (1996). A study of students' Anxiety over Classroom English (6). *Bulletin of Joetsu University of Education*, 15(2), 495–506. <https://juen.repo.nii.ac.jp/records/4860>
- Kaplan-Rakowski, R., & Gruber, A. (2023). The impact of high-immersion virtual reality on foreign language anxiety. *Smart Learn. Environn.*, 10, 46. <https://doi.org/10.1186/s40561-023-00263-9>
- Krashen, S. (1982). *Principles and practice in second language acquisition*. Pergamon Press. https://www.sdkrashen.com/content/books/principles_and_practice.pdf

- Li, M., & Yu, Z. (2023). A systematic review on the metaverse-based blended English learning. *Frontiers in Psychology*, 13. <https://doi.org/10.3389/fpsyg.2022.1087508>
- Li, X., Ishi, C. T., & Hayashi, R. (2021). Analyses of voice quality and acoustic features in Japanese and Chinese emotional speech: Japanese native speakers and Mandarin Chinese learners. *Journal of the Phonetic Society of Japan*, 25, 9–22. https://doi.org/10.24467/onseikenkyu.25.0_9
- Mehrabian, A. (1971). *Silent messages*. Wadsworth Publishing Company.
- Takahashi, E. (2024, August 29). English conversation lessons for Japanese EFL students using the metaverse [Conference presentation]. EUROCALL 2024, Trnava University, Slovakia.
- Takano, Y., & Noda, A. (1995). Interlanguage dissimilarity enhances the decline of thinking ability during foreign language processing. *Language Learning*, 45(4), 657–681. <https://doi.org/10.1111/j.1467-1770.1995.tb00457.x>
- Yee, N., Bailenson, J. N., & Ducheneaut, N. (2009). The Proteus effect: Implications of transformed digital self-representation on online and offline behavior. *Communication Research*, 36(2), 285–312. <https://doi.org/10.1177/0093650208330254>