Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Potential of Synthetic Voice in Multilingual Educational Environments: Utilizing Generative AI in XR and Metaverse-Based International Virtual Exchange

Marvin EDER[a*], **Masako HAYASHI**[a]
[a]*Tohoku University, JAPAN*
*eder.marvin.p2@dc.tohoku.ac.jp

Abstract: In this work-in-progress paper, we analyze the impact of using synthetic, natural, and text-to-speech (TTS) voices in a flipped classroom-style lecture video setting. Our study aims to give an insight on how different voice types affect students' comprehension, perception of naturalness, preferences and overall satisfaction, especially in an international and multilingual learning environment. Participants watched three versions of the same lecture, each only differing in the voice used for narration and evaluated them across various criteria. Results showed that while a natural voice remains slightly preferred, high-quality synthetic voices can perform comparability in terms of clarity and acceptability.

Keywords: Flipped Classroom, Generative AI, Synthetic Voice, International Virtual Exchange, XR-based Education, Multilingual Environment

1. Introduction

As flipped classroom models continue to reshape higher education, video-based lectures have emerged as a central component of asynchronous, self-paced learning. These formats empower students to engage with instructional content on their own schedule, increasing flexibility and accessibility, two widely recognized benefits of flipped learning environments (Akçayır G. & Akçayır M., 2018).

This approach is especially valuable in XR- and metaverse-based international courses, where real-time sessions demand high cognitive effort from students and significant preparation from instructors. Flipped learning minimizes in-class inefficiencies by offloading foundational content to pre-recorded lectures, allowing synchronous time to focus on collaboration, interaction, and experiential learning.

A crucial element of such video lectures is the narration voice, which significantly affects learners' comprehension, engagement, and perceived instructional quality. Research shows that voice characteristics, such as naturalness, human-likeness, and intonation, play a critical role in both the clarity of delivered content and the social presence of the speaker (Kühne et al., 2020). These qualities can foster more immersive and motivating learning environments.

Recent advances in AI-based voice synthesis, including techniques such as few-shot voice cloning and zero-shot Text-to-Speech (TTS), now allow for the rapid generation of high-quality, natural-sounding synthetic speech. These developments enable cost-effective and scalable alternatives to traditional human narration, especially in contexts where personalization and multilingual delivery are desirable (Azzuni & El Saddik, 2025).

This study investigates the impact of narration voice on learners' perception and comprehension through a comparative analysis of three versions of the same flipped lecture: one narrated by a human voice, one by an advanced AI-cloned voice, and one using conventional TTS synthesis.

2. Methods

This section outlines the tools and methods used in the first experiment. It includes the TTS-based video creation system (2.1) and the experimental design and evaluation procedure (2.2).

2.1 Video Creation Tool using GenAI and TTS

A self-developed, cross-platform video generation pipeline was used to transform PowerPoint-based lectures into bilingual narrated videos. The system integrates Google Cloud Text-To-Speech, using *JP-Chirp3* (Japanese) and *en-US-Studio* (English) voice models. Python libraries such as *python-pptx*, *moviepy*, and *pydub* handle slide processing, video rendering, and audio chunking, respectively.

Beyond standard TTS models, a custom-trained synthetic voice was developed using the open-source *Applio* framework. This allowed for the generation of a highly personalized voice, replicating the speaker's pitch, intonation, and rhythm. The model supports multilingual output while preserving the original vocal characteristics.

2.2 Research Methods

Alongside the two GenAI-generated voices, a professionally recorded human voice by the course instructor (native Japanese speaking) was included as a baseline. The participant group (32 students, both Japanese and international) was divided into three groups, further subdivided into three teams for each group, each watching the same three videos (TTS, Trained, Natural) in a different order to counterbalance potential sequence effects.

English videos were shown to English-proficient students (Group 1), and Japanese versions to those more fluent in Japanese (Group 2 and Group 3). After viewing, students completed a standardized questionnaire rating each voice on a 5-point Likert scale across four criteria:

1. **Naturalness** – perceived naturalness of the voice
2. **Human-likeness** – perceived human-likeness of the voice
3. **Understandability** – ease of comprehension
4. **Preference** – willingness to hear the voice in future lectures

3. Results

Figure 1 represents the analysis of the average user rating for the three types of voices - TTS-generated, trained synthetic voice, and natural human voice - across the above-mentioned evaluation categories: Naturalness of the voice (Natural), Human-likeness of the voice (Human-like), Understandability (Understandable), and Prefer to the voice listen again (Prefer).

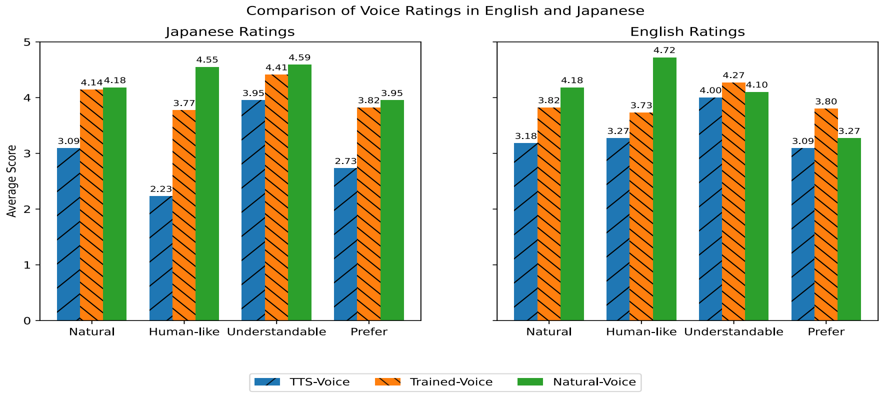

Figure 1. Comparison of Voice Evaluation Scores Across Languages and Voice Types

In addition, the rating is split by language in accordance with the separate evaluation depending on language proficiency.

In the English part of the ratings, the Natural-Voice outperformed both the TTS-Voice and Trained-Voice in both categories related to speech naturalness, achieving scores of 4.72 in the Human-like category, compared to scores of 3.27 for the TTS-Voice and 3.73 for Trained-Voice, and 4.18 in the Natural evaluation, compared to 3.18 (TTS-Voice) and 3.82 (Trained-Voice). The AI voices excel in the other two categories - Understandability and Prefer to listen again, with the Trained-Voice outperforming the other options in both categories, especially when looking at the Prefer rating (3.8) compared to the Natural-Voice (3.27).

In contrast, the Japanese ratings show a more positively defined reception of the *Trained-* and *Natural-Voice*, while the *TTS-Voice* lagged notably in the categories *Human-Like* (2.23) and *Natural* (3.09), indicating limitations in expressiveness and pronunciation for Japanese when comparing it to the *Human-Voice* of a native Japanese speaker. The *Trained-Voice* and *Natural-Voice* received comparable ratings in all categories, with *Human-Like* being the only exception where the *Trained-Voice* significantly falls behind (3.77 vs 4.55) but still performed way better than the *TTS-Voice*.

When comparing the average scores across all rating categories for all voices in both languages, it is noticeable that the *Natural-Voice* overall remains the best-rated option, with a score of 4.19. It is closely followed by the *Trained-Voice*, scoring 3.97, while the *TTS-Voice* falls behind with an average rating of only 3.19. Indicating that a well-trained synthetic voice can closely match or even sometimes surpass a *Natural-Voice*, especially when looking at language localization in a language that is not native to the *Natural-Voice*.

4. Discussion and Conclusion

Through the first results obtained by a small group of participants, we can confirm that a natural narration voice, in their native language, is still the most well-received type of voice overall but closely followed by the custom-trained synthetic voice model. Especially when talking about the localization into a foreign language, the synthetic voice can even outperform the other voice types in categories such as *Ease of understanding* and *Willingness to listen again.* The generic TTS voice models were overall perceived as the worst-performing narration option, falling significantly behind in the *Speech naturalness* metrics, *Natural* and *Human-like.*

Regarding the results for the English-localized lecture videos, it is important to note that the sample size was limited to only 10 students, compared to 22 students in the Japanese reference group. As a result, the findings may not be fully generalizable to the broader student population and should be interpreted with caution. This highlights the need to replicate the experiment with a larger and more diverse sample group.

Acknowledgements

This work was supported in part by JSPS KAKENHI Grants JP23KK0037 and JP24K06075, The Open University of Japan, and Tohoku University Advanced Institute of So-Go-Chi Informatics. We thank VRChat, Inc. for providing the metaverse environment.

References

Azzuni, H., & El Saddik, A. (2025). Voice Cloning: Comprehensive Survey. *ArXiv.org*. https://arxiv.org/abs/2505.00579

Akçayır, G., & Akçayır, M. (2018). The flipped classroom: A review of its advantages and challenges. *Computers & Education, 126(1),* 334–345. https://doi.org/10.1016/j.compedu.2018.07.021

Kühne, K., Fischer, M. H., & Zhou, Y. (2020). The Human Takes It All: Humanlike Synthesized Voices Are Perceived as Less Eerie and More Likable. Evidence From a Subjective Ratings Study. *Frontiers in Neurorobotics, 14.* https://doi.org/10.3389/fnbot.2020.593732