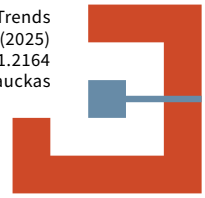
Trends

JALTCALL Trends
ISSN 2982-3749, 1, 2164 (2025)
https://doi.org/10.29140/jct.v1n1.2164



Castledown

Learner interactions with Jouzu, the mobile application with conversational characters

Zackary Rackauckas
Columbia University
zcr2105@columbia.edu

Jouzu (https://www.jouzu.ai/) is a mobile app that supports Japanese learners through immersive, judgment-free conversations with fictional characters. By integrating large language models (LLMs) and text-to-speech (TTS) synthesis, Jouzu brings these characters to life with distinct speech patterns inspired by Japanese pop culture. The app enables spoken and written interaction tailored to different proficiency levels. This paper examines how learners engage with these characters and how interactive, voiced dialogue enhances vocabulary acquisition and conversational fluency beyond traditional language learning methods.

Keywords: AI in education, computer-assisted language learning, chatbot

Introduction

Japanese is ranked among the most difficult languages for native English speakers to learn, requiring approximately 2,200 classroom hours to reach proficiency (US Department of State, 2024). Beyond its multiple writing systems, Japanese diverges significantly from English in grammar (Boboc, 2023) and phonology (Ohata, 2004). Traditional tools such as textbooks are often rigid, impersonal, and lack spoken interaction. These factors contribute to learner disengagement and high dropout rates.

Conversational practice remains one of the most effective techniques for language acquisition, but it is often limited by social anxiety and access to native speakers (Li et al., 2023). Recent advancements in computer-assisted language learning (CALL) have introduced novel solutions. Researchers have used large language models (LLMs) to support automated lesson planning (Dornburg, 2024) and generate student feedback (Cao, 2023). Some have designed question-answer (QA) chatbots for simulated conversation (Li et al., 2022), and others have incorporated empathetic responses to reduce anxiety in language learning (Siyan et al., 2024a; 2024b). Aiba et al. (2024) further demonstrated the potential of LLMs by pairing ChatGPT with TTS to simulate academic conferences for English learners.

Jouzu builds on such work by simulating fictional characters with distinct

character language (Kinsui & Yamakido, 2015), enabling immersive and emotionally engaging Japanese practice. These conversational characters offer learners multimodal, conversational environments – typed or spoken – that replicate human interaction without judgment.

Character implementation

Character creation

Three fictional characters were created in the spirit of Japanese anime and manga: Hikari Majo (Figure 1), Kitsune Jouou (Figure 2), and Saya Musume (Figure 3). These characters each exhibit their own character language based on established character tropes.

Figure 1. Hikari Majo

Figure 2. Kitsune Jouou

Figure 3. Saya Musume

Hikari, an energetic magic school student, uses casual speech patterns. She frequently uses the emphatic phrase *noda*, a marker of youth and excitement.

> 魔法の森で迷子になっちゃって、光魔法でサインを送ったら、妖精が助けてくれたんだよ。めっちゃドキドキしたけど、すっごいキレイな光景が見れて、最高の冒険だったの！
>
> I got lost in the magical forest, so I used light magic to send a signal, and a fairy came to my rescue! I was really nervous, but the scenery was so beautiful – it turned out to be the best adventure ever!

Kitsune, a centuries-old fox deity, uses archaic and poetic phrasing. She prefers the first-person pronoun *warawa*, the copula *ja*, and the classical aspect form *-teoru*, instead of modern *-teiru*.

> 秘密は心を静かに保ち、自然のリズムに耳を傾けることにあり、そうすれば、生きとし生けるものとの調和を見出せるというものじゃ。
>
> The secret lies in stilling your heart and attuning yourself to the gentle rhythm of nature. In that quiet harmony, you shall discover the balance with all living things.

Saya, a tomboyish samurai, speaks with formal, masculine samurai language. She uses the first-person pronoun *sessha*, the archaic, formal copula *degozaru*, and the emphatic particle *zo*.

拙者がしっかりと指導致すので、焦らずに根気強く練習を積むでござる。精進あるのみでござるぞ。

I shall instruct you with utmost diligence. Do not hasten; instead, train with patience and unwavering perseverance. The path to mastery lies in relentless dedication.

These characters were selected to expose learners to diverse registers of Japanese – from casual modern speech to poetic archaic forms – broadening their listening comprehension. The speech patterns were inspired by common tropes in anime, where female and non-standard dialects are often associated with expressive or distinctive personalities.

Data

Using OpenAI's GPT-4, long-form QA scripts were generated between users and each character. These scripts were reviewed and corrected by a native speaker for natural fluency. Professional Japanese voice actors recorded the finalized scripts, adding tone and personality to each character.

Training

The variational inference text-to-speech (VITS) framework generates audio that nearly matches human speech as shown by comparative mean opinion score (Kim, 2021). Using the recorded character dialogue as training data, the vits-uma-genshin-honkai (zhomehwh, 2024) pretrained Chinese and Japanese multilingual VITS model was fine-tuned using VITS-fast-finetuning (Plachtaa et al, 2024). The three fine-tuned models were deployed on a web application programming interface (API) using vits-simple-api (Artrajz et al., 2024).

User experience

The user experience of the conversational character system is composed of the following steps:

- The user types or speaks a message to the character.
- The character responds with text generated by the LLM and speech audio synthesized by the VITS model.
- The user continues the conversation chain with each message stored in conversational history.

The backend process consists of the following steps:

- The text input or transcribed audio input is received.
- The LLM is given the user input and customized prompt to generate character language.
- The VITS model synthesizes the audio.
- Text and audio are fused into the character's output.

Learner interaction

Beginner usage

Preliminary user interactions from a user study with 32 participants reveal three interaction patterns with the chat feature of Jouzu. First, several users used English to ask Hikari to "teach [them] a word in Japanese." Then, characters responded with a Japanese sentence explaining a simple word.

> **User**: Teach me a word in Japanese.
> **Hikari**: うん、いいよ!「こんにちは」って言うの。これは「Hello」だよ。
> (Translation) Sure! You can say "Hello." That's "こんにちは" in Japanese.

Next, users asked how to say certain words in Japanese. This is a common approach beginners use when learning a new language (Rodríguez Cervantes et al., 2012).

> **User**: How do you say sword in Japanese?
> **Saya**:「刀(かたな)」と申すでござる。
> (Translation) You say "katana."

Third, some users engaged in basic conversations with characters using English input. The characters answered in Japanese, providing an immersive conversational experience with a low barrier to entry, since users did not have to know Japanese.

> **User**: What is your favorite tea?
> **Hikari**: 光は、甘いミルクティーが好きよ。
> (Translation) I really like sweet milk tea.

These dialogue patterns often continued consistently until the learner exited the session.

Advanced usage

Students with prior knowledge of Japanese used the conversational characters to simulate real Japanese conversations. Similar to beginners, conversations centered around topics such as preferences, but more advanced users engaged using spoken Japanese through the app's speech recognition system. Based on these interactions, it is hypothesized that Jouzu offers a low-pressure, near-native conversation environment where learners can acquire new vocabulary and practice speaking in context. Future analysis of user engagement and survey data will further evaluate this hypothesis.

Discussion

Practical insights

At the beginning of development, OpenAI GPT models were selected primarily due to the ease of integration with the API. At that time, GPT-4 was OpenAI's flagship model, delivering near-human performance across various tasks. Although GPT-4 generated highly accurate and natural character language, each response required

approximately 20 seconds to generate. In May 2024, OpenAI introduced GPT-4o, a model that surpassed previous benchmarks and showcased impressive multimodal capabilities (Ying, 2024). GPT-4o did not immediately achieve the same character dialogue quality with identical prompts, but by providing more conversational context, similar results were attained. The switch to GPT-4o also reduced response times to 1-2 seconds and cut resource overhead by 50 percent. VITS was chosen for high-quality speech synthesis; however, other multilingual models such as Style-BERT-VITS2 (litagin02, 2023), GPT-SoVITS (RVC-Boss, 2023), and Emotional-VITS (innnky, 2023) were also considered but not implemented. Further experimentation with these models remains an area for future work.

Bias and sensitivity

To prevent potential biases in character outputs, several safeguards were implemented to filter inappropriate responses and ensure that the dialogue remained culturally sensitive and educationally appropriate (Xue et al., 2023). Prompts and responses were monitored throughout development to minimize the likelihood of biased or harmful content. These measures were taken to ensure that the generated dialogue aligns with educational goals and promotes a positive learning environment (Chan, 2023).

Conclusion

Jouzu was developed to support Japanese language learners through multimodal, interactive dialogue. Conversational characters operate through the integration of LLMs and TTS, offering spoken and written interaction. Multimodal input is supported through speech recognition or typed queries. Future research includes evaluating new language models and conducting iterative user testing to refine and improve the learning experience. As large language models and speech synthesis continue to improve, character-driven systems like Jouzu will become integral tools for supporting immersive, low-anxiety language learning across diverse contexts.

Acknowledgements

This research was initiated independently of the author's academic responsibilities at Columbia University and was conducted without the use of university resources.

References

Aiba, M., Saito, D., & Minematsu, N. (2024, September). A ChatGPT-based oral Q&A practice system for first-time student participants in international conferences. In *Proceedings of Interspeech 2024* (pp. 5202–5203). ISCA. https://www.isca-archive.org/interspeech_2024/aiba24_interspeech.pdf

Artrajz, Litss, D., & Risenafis, et al. (2024, August 25). *vits-simple-api* [Computer software]. GitHub. https://github.com/Artrajz/vits-simple-api

Boboc, V. (2023). An algebraic approach to translating Japanese. *Journal of Language Modelling, 11*(1). https://doi.org/10.15398/jlm.v11i1.315

Cao, S., & Zhong, L. (2023). Exploring the effectiveness of ChatGPT-based feedback compared with teacher feedback and self-feedback: Evidence from Chinese to English translation. *arXiv*. https://arxiv.org/abs/2309.01645

Chan, C. K. Y., & Tsi, L. H. Y. (2023). The AI revolution in education: Will AI replace or assist teachers in higher education? *arXiv*. https://arxiv.org/abs/2305.01185

Dornburg, A., & Davin, K. (2024). To what extent is ChatGPT useful for language teacher lesson plan creation? *arXiv*. https://arxiv.org/abs/2407.09974

innnky. (2023). *Emotional-VITS* [Computer software]. GitHub. https://github.com/innnky/emotional-vits

Kim, J., Kong, J., & Son, J. (2021). Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. *arXiv*. https://arxiv.org/abs/2106.06103

Kinsui, S., & Yamakido, H. (2015). Role language and character language. *Acta Linguistica Asiatica, 5*(2), 29–42. https://doi.org/10.4312/ala.5.2.29-42

Li, C., Leng, Z., Yan, C., & Yu, D., et al. (2023). ChatHaruhi: Reviving anime character in reality via large language model. *arXiv*. https://arxiv.org/abs/2308.09597

Li, Y., Chen, C.-Y., Yu, D., et al. (2022). Using chatbots to teach languages. In *Proceedings of the Ninth ACM Conference on Learning @ Scale*. https://doi.org/10.1145/3491140.3528329

litagin02. (2023). *Style-BERT-VITS2* [Computer software]. GitHub. https://github.com/litagin02/Style-Bert-VITS2

Siyan, L., Shao, T., Yu, Z., & Hirschberg, J. (2024a). EDEN: Empathetic dialogues for English learning. *arXiv*. https://arxiv.org/abs/2406.17982

Siyan, L., Shao, T., Yu, Z., & Hirschberg, J. (2024b). Using adaptive empathetic responses for teaching English. *arXiv*. https://arxiv.org/abs/2404.13764

Ohata, K. (2004). Phonological differences between Japanese and English: Several potentially problematic areas of pronunciation for Japanese ESL/EFL learners. *Asian EFL Journal, 6*(1).

Plachtaa, A. K., Artrajz, John, J., & lrioxh, et al. (2024, July 3). *VITS-fast-fine-tuning* [Computer software]. GitHub. https://github.com/Plachtaa/VITS-fast-fine-tuning

Rodríguez Cervantes, C. A., & Roux Rodriguez, R. (2012, November). The use of communication strategies in the beginner EFL classroom. *GIST Education and Learning Research Journal, (6)*, 111–128. https://eric.ed.gov/?id=EJ1062702

RVC-Boss. (2023). *GPT-SoVITS* [Computer software]. GitHub. https://github.com/RVC-Boss/GPT-SoVITS

US Department of State. (2024). Foreign language training. Retrieved August 24, 2024, from https://www.state.gov/foreign-language-training

Xue, J., Wang, Y.-C., Wei, C., & Liu, D. (2023). Bias and fairness in chatbots: An overview. arXiv. https://arxiv.org/abs/2309.08836

Ying, Z., Liu, A., Liu, X., & Tao, D. (2024). Unveiling the safety of GPT-4o: An empirical study using jailbreak attacks. *arXiv*. https://arxiv.org/abs/2406.06302

zomehwh. (2024). *VITS-Umamusume, Genshin, Honkai voice synthesizer* [Computer software]. Hugging Face. https://huggingface.co/spaces/zomehwh/vits-uma-genshin-honkai

Author biodata

Zackary Rackauckas is a student at Columbia University on the Master of Science, Computer Science Thesis track advised by Dr. Julia Hirschberg. His research interest is leveraging natural language processing and speech processing for computer-assisted language learning, especially for the English and Japanese languages.