Computers & Graphics 133 (2025) 104473

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Special Section on ICXR 2025

Voice of artifacts: Evaluating user preferences for artifact voice in VR museums☆

Bingqing Chen [a], Wenqi Chu [a], Xubo Yang [b], Yue Li [a,*]

[a] School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou, China
[b] School of Software, Shanghai Jiao Tong University, Shanghai, China

ARTICLE INFO

Keywords:
Virtual museum
Museum chatbots
Text-to-Speech Models (TTS)
Interactive vocal experiences
User preferences

ABSTRACT

Voice is a powerful medium for conveying personality, emotion, and social presence, yet its role in cultural contexts such as virtual museums remains underexplored. While prior research in virtual reality (VR) has focused on ambient soundscapes or system-driven narration, little is known about what kinds of artifact voices users actually prefer, or if customized voices influence their experience. In this study, we designed a virtual museum and examined user perceptions of three types of voices for artifact chatbots, including a neutral synthetic voice (*default*), a socially relatable voice (*familiar*), and a user-customized voice with adjustable elements (*customized*). Through a within-subjects experiment, we measured user experience with established scales and a semi-structured interview. Results showed a strong user preference for the *customized* voice, which significantly outperformed the other two conditions. These findings suggest that users not only expect artifacts to speak, but also prefer to have control over the voices, which can enhance their experience and engagement. Our findings provide empirical evidence for the importance of voice customization in virtual museums and lay the groundwork for future design of interactive, user-centered sound and vocal experiences in VR environments.

1. Introduction

Museums are not silent spaces - they are rich, multisensory environments where sound plays a key role in shaping visitor experience. From the ambient noise of footsteps and murmurs to curated audio guides and interactive exhibits, museums are filled with diverse auditory elements [1,2]. Among these, human voice holds a particularly powerful position, which can guide, narrate, question, and emotionally connect with visitors [3]. As museums evolve to offer more immersive and interactive experiences, voice-based interactions are becoming an increasingly important component of digital engagement strategies. The use of chatbot voices in virtual reality (VR) museums has shown great potential to enhance the visiting experience by providing personalized guidance, narration, and emotional resonance [4,5]. While prior work has explored the roles these agents can take, such as docents, historical figures, or creators [6], relatively few studies have examined how different types of voices themselves shape user experience. For instance, Garcia et al. [7] used a humanoid conversational agent with a default synthetic voice to guide visitors through a museum, and Noh and Hong [8] designed speaking artifacts using pre-recorded narration to enrich visitor interpretation. Despite these advances, the design and customization of chatbot voices in cultural settings remain underexplored. Specifically, the role of voice in museums remains underexplored, especially in terms of how different vocal styles and levels of customization influence user experience. This study aims to address that gap by investigating user preferences for various chatbot voices and examining whether vocal customization affects visitor experience in a virtual museum setting. Therefore, our guiding research questions are,

RQ1 What type of chatbot voices do users prefer for the museum artifact chatbot?
RQ2 Does the degree of customization in artifact voices affect visitor experience?

We conducted a user study with 21 participants to evaluate preferences for three types of chatbot voices in a VR museum context. Results revealed a clear preference for voices with high-level customization, which were appreciated for their varied timbres and the ability to be adapted to different artifacts' appearances, thereby enhancing

☆ This article is part of a Special issue entitled: 'CAG_ICXR 2025' published in Computers & Graphics.

* Corresponding author.
E-mail addresses: bingqing.chen16@student.xjtlu.edu.cn (B. Chen), wenqi.chu22@student.xjtlu.edu.cn (W. Chu), yangxubo@sjtu.edu.cn (X. Yang), yue.li@xjtlu.edu.cn (Y. Li).

https://doi.org/10.1016/j.cag.2025.104473
Received 19 October 2025; Received in revised form 30 October 2025; Accepted 2 November 2025
Available online 8 November 2025

Fig. 1. Six artifacts in the virtual museum. (a) Harbin Institute of Technology Commemorative Plate; (b) Elephant Ornament; (c) MingZhou Yue Kiln Celadon; (d) Confucius Statue; (e) Column of Tongji University; (f) Ceramic Jar.

the sense of engagement. However, participants also noted concerns about the longer voice generation time and the lack of detailed description of each adjustable dimension. These findings offer valuable insights for the design of voice-based interaction in cultural heritage settings, highlighting the importance of customization in shaping user engagement.

2. Related work

2.1. Voice in virtual museums

Voice is a specific type of sound, produced by the human vocal tract when speaking, singing, or making other vocal noises. It is inherently interpersonal and interactive, enabling the conveyance of narrative, character, and presence in both physical and virtual spaces [9]. Recent studies have highlighted that the integration of voice in virtual environments significantly alters users' sense of presence and social engagement [10].

Voice can originate from various sources, including real users via voice chat [10] and virtual agents such as guides [11], assistants [12], and embodied characters [13]. For example, van Rijn et al. [14] developed a system allowing users to give a voice to a robot and adjust its sound based on visual appearance, enhancing engagement during interaction. Within virtual museums, voice is typically associated with avatars representing digital docents [15,16]. These voices are often used to deliver narration, provide contextual information, or simulate conversations [17,18], enriching user engagement and interpretive depth.

Researchers have explored how voices with a specific identity affect users' experiences. For example, Al-Taie et al. [19] investigated five categories of soundscapes in a PC-based virtual museum, including recorded human voices. Their study found that soundscapes tailored to the exhibited objects significantly improved users' engagement, suggesting that context-specific auditory elements can enhance the interpretive experience in virtual exhibitions. Also, Salo et al. [20] developed a system that allowed users to record and upload their own stories related to museum artifacts. Their findings suggest that enabling visitors to contribute personal audio narratives fosters a sense of participation and engagement, supporting more meaningful and emotionally resonant museum experiences.

While these existing research studies are effective in providing contextual information, most voice implementations in virtual museums are system-driven and non-customized, offering the same audio experience with the same voice for different scenarios to all users. This approach fails to accommodate individual preferences. There is growing interest in exploring how user-tailored voices could enrich cultural engagement. Yet few studies have investigated how users respond to different types of artifact voices, especially those that vary in customization or expressive control. Given the unique potential of voice to embody character, emotion, and cultural context, this work advocates for a shift from curated auditory design toward user-centered and customizable audio experiences in VR museums. Such a shift reconceptualizes voice not only as a delivery channel, but as a co-creative medium—capable of fostering connection in digitally mediated heritage spaces.

2.2. Chatbots with vocal expression

Recent advances in natural language generation and multimodal integration have enabled users to engage in meaningful, open-ended interactions with conversational agents. For instance, Qin et al. [21] present CharacterMeet, a GPT-4–powered character development tool that supports exploratory, meaning-making workflows rather than fixed scripts. In addition, existing text-to-speech (TTS) models are widely used as the voices of various chatbots, including physical [22,23] and virtual ones [10]. Their quality has improved significantly over the past decade [24,25], with many now capable of producing voices that are nearly indistinguishable from human recordings [26].

In the context of museums, conversational agents with voice interaction capabilities have been shown to enhance visitor engagement by providing interactive, entertaining, and consistently available alternatives to traditional guided tours. For example, Trichopoulos et al. [27] combined ChatGPT-4 and Whisper to create a personalized and immersive museum guide, while Duguleană et al. [18] designed a virtual tour of the Casa Muresenilor Museum that lets users explore artifacts interactively. These agents can take on various roles, including traditional docent-style guides [28], historical figures [29,30], artifacts [8], or even the creators of the artifacts themselves [6]. For example, Garcia et al. [7] developed a stylized humanoid conversational agent to act as a museum guide, utilizing the default text-to-speech voice provided by ChatGPT-4, enhancing user engagement and interest. Trovato et al. [31] implemented a voice-based robot guide for Q&A in a museum setting using a default female voice generated by a standard TTS engine, exploring the categories users often asked. While many designers attempted to give voice to museum chatbots, the focus has remained largely on the content of what the agent delivers, rather than the nature of its voice or who is perceived to be speaking.

As a result, this study aims to investigate how to design more immersive chatbot voices that encourage users to engage more deeply with virtual cultural artifacts. Building on the findings of previous studies [6,14,20], we introduce three types of voices of the artifact chatbot for comparison: the *default*, *familiar*, and *customized*. This study seeks to explore user preferences across these voice types, laying the groundwork for future design in voice-based interaction within cultural heritage contexts and VR environments.

3. Methodology

To explore user preferences for artifact voice types in virtual museums, we designed a virtual museum system and conducted evaluation studies with participants.

3.1. System design and implementation

3.1.1. Modeling of artifacts

We obtained 3D models of six artifacts from a local museum, as shown in Fig. 1. They were scanned using an iPhone 16 Pro and reconstructed using Vuforia Creator.[1] The detailed artifact information is provided in Appendix A.

[1] https://apps.apple.com/us/app/vuforia-creator/id1625877749

Table 1
Features of three types of chatbots.

Voice type	Description	User control
Default	A default, system-generated voice with a neutral tone.	Gender: male, female
Familiar	A voice mimicking someone the user knows, such as a close friend.	Speaker identity
Customized	A user-defined voice that can be adjusted in different dimensions.	Sound elements

3.1.2. Chatbot voice types

Based on prior research [6,14,20], we designed three types of voices for the artifact chatbot, with their characteristics summarized in Table 1.

1. For the *default* voice, we emulated the type of voice commonly used in audio guide devices found in physical museums, offering a low degree of customization. Two options were provided for users to choose from, including a male and a female voice.
2. The design of the *familiar* voice is inspired by the approaches of Salo et al. [20]. The system allowed users to record the voice of someone familiar, such as themselves or a friend, to serve as the basis for speech synthesis. Users are not allowed to edit any specific features of the voice once it is synthesized.
3. The *customized* voice allowed a greater degree of control. Fig. 2 illustrates the implementation of the *customized* voice tool. The tool enables users to customize artifact voices through a visual interface powered by a Variational Inference with adversarial learning for an end-to-end Text-to-Speech model (VITS) trained on a Mandarin dataset. It utilizes speaker embeddings ($n = 186$), which are reduced via Principal Component Analysis (PCA) and mapped to a 5-dimensional voice configuration tool. Each dimension was presented using a slider ranging from 0 to 1. We omit displaying descriptive labels (e.g., "male–female", "neutral–expressive") because the axes are linear combinations of acoustic/phonetic cues and does not have a stable, one-to-one human-interpretable meaning [14]. Instead, participants freely explored the five sliders, auditioned the synthesized voice, and confirmed a preferred setting. When they reach a setting they liked, the slider values are mapped back into the 5-D space to reconstruct a speaker embedding. This embedding, along with input text, is passed to the TTS model to generate artifact-specific speech. The resulting waveform forms the final output, allowing users to match different voices to different artifacts' appearances.

3.1.3. Virtual museum

A virtual museum with 6 different artifacts was built in Unity 2022.3.44f1c1 (see Fig. 3b). Each artifact is accompanied by a panel with textual descriptions, similar to the settings in a physical museum. This information was sourced from the museum's official website. In addition, the artifact 'speaks' through a first-person narrative, generated from its official exhibition description. All narratives were reviewed by a museum researcher to ensure accuracy and authenticity. Along with the textual descriptions, the panel also presents three voice options. Users can switch between them by clicking the corresponding buttons. The interface for *default* voice is shown in Fig. 3c, where users can hear the voice by simply clicking the *Play* button. We deployed the FishSpeech TTS model locally.[2] The *default* voices are generated by directly inputting the textual description of each artifact. For the *familiar* voice, users can either read a prepared text or enter a customized text script into the *Text during recording* input box. They can then click *Record Sound* to begin recording. When they click *Play*, the system will play a first-person narration using their recorded voice (see Fig. 3d). We adopted the reference function of the FishSpeech TTS model, which takes the recorded voice as a reference and synthesizes the same voice based on the input. The *customized* voice interface is shown in Fig. 3e. Users can adjust the sliders (see Fig. 3f) and then click *Play* to generate a customized voice. If they are unsure where to start, a *Random* button allows them to explore different voice variations. We adopted the VITS Chinese TTS model[3] and a Mandarin dataset[4] in the training.

3.2. Measures

We assessed visitors' experience using three validated instruments for museum experience, multimedia guide, and user experience. Specifically, the Museum Experience Scale (MES) [32] captures engagement, knowledge and learning, meaningful experience, and emotional connection on a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree). In addition, we adopted the Multimedia Museum Guide Scale (MMGS) [32] to measure general usability, learnability & control, and quality of interaction. These were measured on the same 1–5 scale. Furthermore, the short form of the User Experience Questionnaire (UEQ-S) [33] was used to evaluate the pragmatic and hedonic qualities of the system. This questionnaire was rated on bipolar items scored from −3 (horrible) to +3 (excellent). All questionnaires were administered after each condition. After completing all conditions, we prepared a follow-up semi-structured interview with open-ended questions, including two questions: (1)Why did you rate the three voice types in this way? and (2) Do you have any other comments or suggestions?

3.3. Experimental settings and procedure

The experiment was conducted in a quiet room (see Fig. 3a). Before commencing the experiment, participants were briefed on the study's purpose, procedure, duration, and data collection methods. After the participants confirmed a clear understanding of the experiment, they were invited to sign an informed consent form. We then collected participants' demographic information, museum visit frequency, and familiarity with the text-to-speech (TTS) and speech-to-text (STT) models. The moderator then helped participants wear the Meta Quest Pro VR HMD to ensure a comfortable fit. The experimental system was implemented on a laptop computer equipped with an NVIDIA RTX 3090 graphics card. The order of the voice system experienced was counterbalanced to mitigate order effects, with an example sequence of the experimental procedure illustrated in Fig. 4. During the experiment, participants were asked to explore each type of voice given to all six artifacts. After finishing experiencing each voice type, we asked the participants to fill in the questionnaires. At the end of the experiment, we asked participants to rate their preference for the three voice types on a 7-point scale, ranging from 1 (strongly dislike) to 7 (strongly like). Finally, a semi-structured interview was conducted.

3.4. Participants

Twenty-one participants (7 females, 14 males) voluntarily signed up for the experiment, with ages ranging from 19 to 26 years old ($M = 21.05$, $SD = 1.80$). More than half ($N = 13$) visited museums once a year or less. On a scale from 1 (not familiar at all) to 5 (very familiar), participants indicated moderate familiarity with text-to-speech (TTS) ($M = 3.14$, $SD = 1.15$) and speech-to-text (STT) technologies ($M = 2.81$, $SD = 1.21$).

[2] https://github.com/fishaudio/fish-speech

[3] https://github.com/Plachtaa/VITS-fast-fine-tuning/blob/main/README_ZH.md

[4] https://github.com/w4123/GenshinVoice

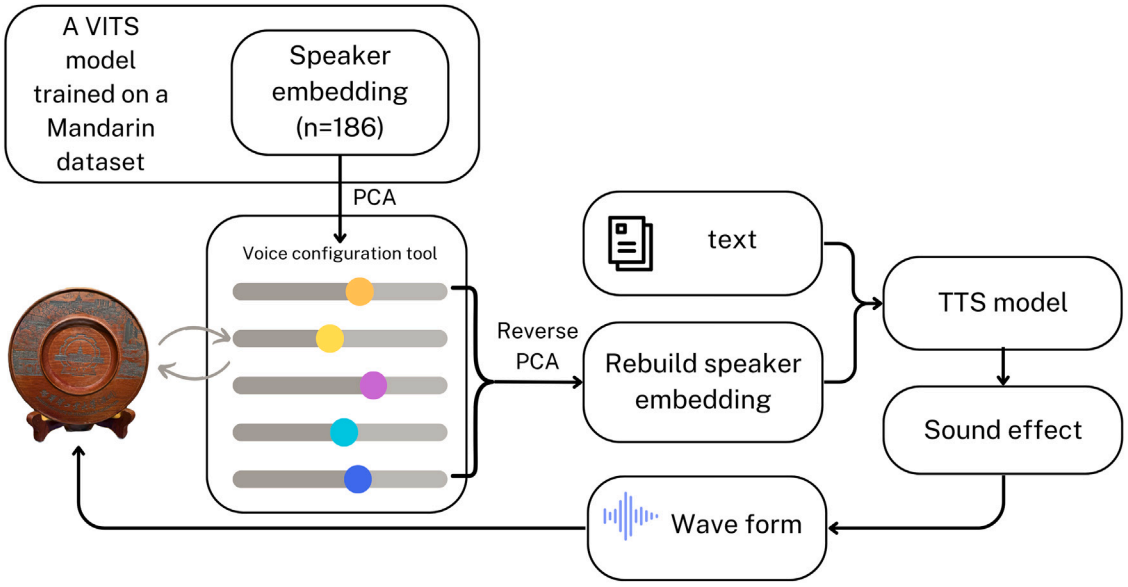

Fig. 2. A diagram illustrating the design and implementation of a customized voice tool.

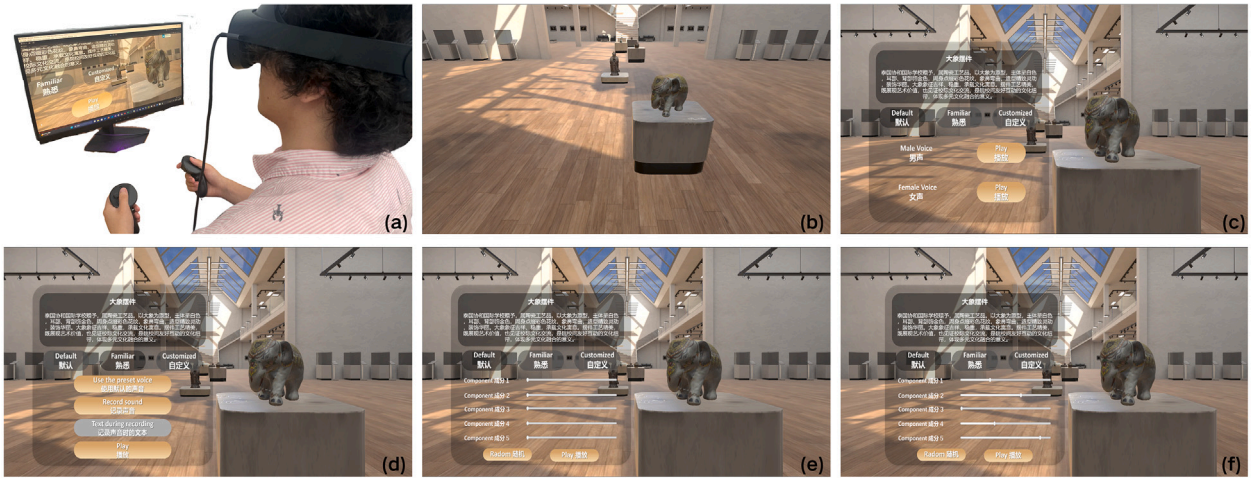

Fig. 3. Illustrations of the virtual museum and three different types of interface for vocal interaction. (a) A participant wearing a VR HMD with two handheld controllers during the experiment; (b) the virtual museum environment with six museum artifacts; (c) an artifact with the introduction panel and the system user interface for *default* voice; (d) the system user interface for *familiar* voice; (e); (f) the system user interface for *customized* voice.

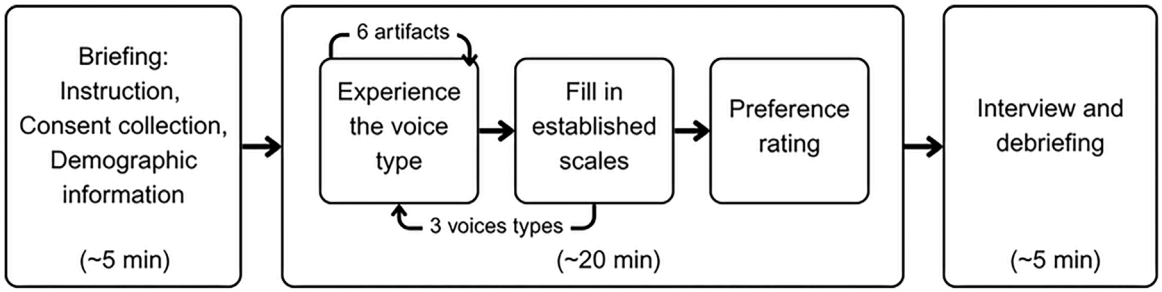

Fig. 4. The experimental procedure for a within-subjects study evaluating the three voice types.

4. Results

We first conducted Shapiro–Wilk tests to assess data distribution. For the comparison across the three voice types, we applied one-way repeated measures ANOVA for normally distributed data and Friedman tests for non-normally distributed data. For qualitative analysis, two researchers coded responses and categorized them into key themes. The coding process involved four predetermined codes: default voice, familiar voice, customized voice, and suggestions. In contrast, specific comments were categorized using an emergent coding approach.

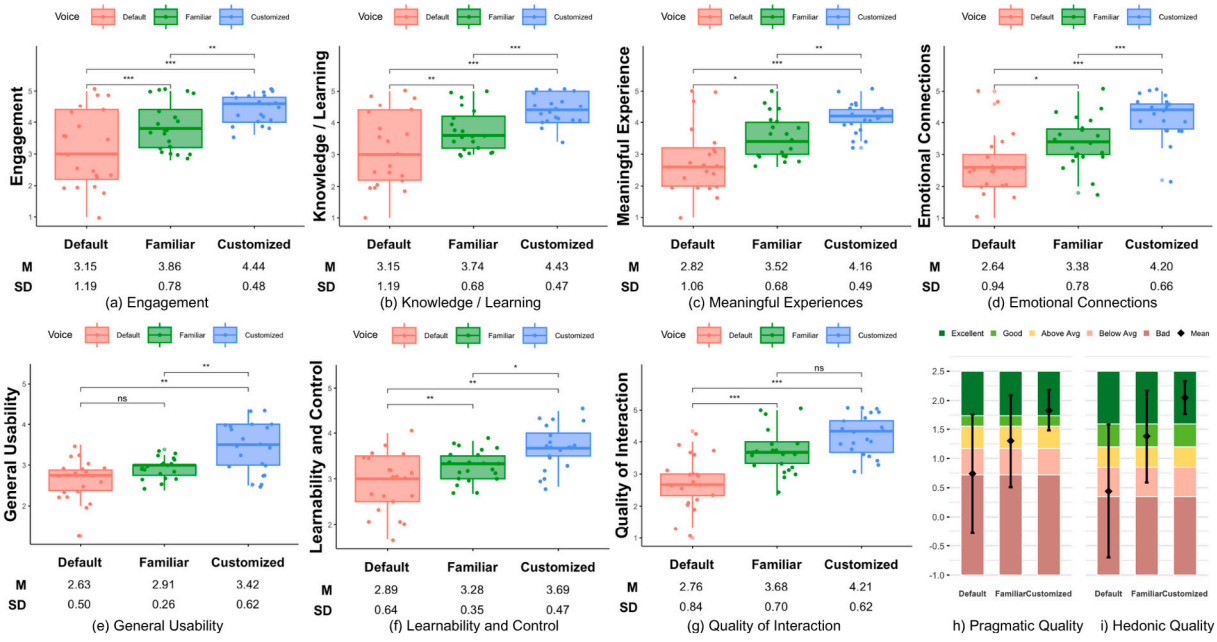

Fig. 5. Box plots and tables of descriptive statistics (means and standard deviations) showing the data analysis results of quantitative analysis: (a) engagement, (b) knowledge/learning, (c) meaningful experience, (d) emotional connection, (e) general usability, (f) learnability and control, (g) quality of interaction, (h) pragmatic quality, (i) hedonic quality. Significance p-value: $^{*}: p < 0.05$, $^{**}: p < 0.01$, $^{***}: p < 0.001$, ns: not significant.

4.1. Quantitative analysis

4.1.1. Museum experience scale

Results for MES are shown in Fig. 5(a-d). Significant differences were found in all subscales. For the *engagement* subscale, a one-way repeated-measures ANOVA was conducted to compare the perceived *engagement* across three voice types, with the analysis revealing a significant effect of voice type ($F(2,40) = 20.03, p < .001, \eta_p^2 = .500$). Post hoc comparisons with Bonferroni correction showed that *customized* voice (M = 4.44, SE = 0.10) was rated significantly higher in engagement than both *familiar* ($p = .007$) and *default* ($p < .001$). In addition, *familiar* voice was rated significantly higher than *default* ($p < .001$). For the *knowledge/learning* subscale, a significant difference was revealed ($F(2,40) = 19.99, p < .001, \eta_p^2 = .500$). Post hoc comparison with Bonferroni correction indicated that the *customized* voice was rated significantly higher than both the *familiar* ($p < .001$) and *default* ($p < .001$) voices. Additionally, the *familiar* voice was rated significantly higher than the *default* voice ($p = .009$). The *meaningful experiences* subscale also showed significant differences ($F(2,40) = 18.59, p < .001, \eta_p^2 = .482$), with *customized* voice being rated significantly higher than both the *familiar* ($p = .003$) and *default* voices ($p < .001$). Additionally, the *familiar* voice was rated significantly higher than the *default* voice ($p = .012$). Finally, the *emotional connection* subscale showed significant effects ($F(2,40) = 26.14, p < .001, \eta_p^2 = .566$), and the *customized* voice was rated significantly higher than both the *familiar* ($p < .001$) and *default* one ($p < .001$). Additionally, the *familiar* voice was rated significantly higher than the *default* voice (p = .013).

4.1.2. Multimedia guide scale

Results for MMGS are illustrated in Fig. 5(e-g). Significant differences were found in all subscales. The analysis for the *general usability* subscale showed a significant difference ($F(2,40) = 11.69, p < .001, \eta_p^2 = .369$), with *customized* voice (M = 3.42, SE = 0.14) rated significantly higher than both the *familiar* ($p = .003$) and *default* voices ($p = .004$). The *learnability and control* subscale also revealed a significant difference ($F(2,40) = 12.42, p < .001, \eta_p^2 = .383$). Post hoc showed that the *customized* voice was rated significantly higher than both the *familiar* ($p = .027$) and *default* voices ($p = .004$). The Familiar voice was also rated significantly higher than the *default* voice ($p = .006$). Similarly, for *quality of interaction* subscale, significant differences ($F(2,40) = 23.20, p < .001, \eta_p^2 = .537$) were observed, with the *customized* voice rated significantly higher than both the *familiar* ($p < .001$) and *default* voices ($p < .001$).

4.1.3. User experience questionnaire

Results for UEQ-S are shown in Fig. 5(h-i). For pragmatic quality, a Friedman test revealed a significant difference in ratings across the three voice conditions ($\chi^2(2) = 20.10, p < .001$). Post-hoc Wilcoxon signed-rank tests with Bonferroni correction indicated that ratings in the *customized* voice condition were significantly higher than both the *default* ($Z = -4.48, p < .001$) and the *familiar* voice conditions ($Z = -2.01, p = .014$). In addition, a Friedman test revealed a significant difference in hedonic quality ratings across the three voice conditions ($\chi^2(2) = 27.79, p < .001$). Post-hoc tests showed that the *customized* voice condition was rated significantly higher than both the *default* ($Z = -4.71, p < .001$) and the *familiar* voice conditions ($Z = -2.47, p = .041$).

4.2. Qualitative analysis

4.2.1. Default voice

Feedback on the *default* Voice condition revealed largely negative responses, which received the lowest average rating ($M = 4.10, SD = 1.48$). A small number of participants (3/21) appreciated the *default* voice for its alignment with the museum context, noting that it resembled a traditional museum guide and supported a formal, informative tone. As P8 and P9 commented, the *default* voice felt appropriate for a museum setting, while P18 added that using a peer's voice might feel out of place and break the sense of presence. However, many participants (13/21) expressed negative impressions, describing the *default* voice as too plain or stiff. These critiques suggest that while the *default* voice maintains contextual appropriateness, it may lack the emotional expressiveness or customization that fosters deeper engagement.

4.2.2. Familiar voice

The *familiar* Voice condition allowed participants to use their own voices or those of their peers, which was generally perceived positively, with an average rating of 5.48 ($SD = 1.29$). Seven participants (7/21) found it interesting and enjoyable, noting that hearing a familiar voice added a sense of presence and novelty to the experience. As P4 and P7 remarked, using one's own or a peer's voice felt refreshing and immersive, while P5 shared that it made the experience more playful and helped them focus better on the content, as if their companion were narrating alongside them. One participant (P5) specifically appreciated hearing their own voice during the interaction. However, a few participants (3/21) noted issues with voice quality or synthesis artifacts. One participant (P14) expressed a sense of dissonance when hearing a peer's voice, suggesting a mismatch between voice familiarity and the formal context of the museum. Overall, the *familiar* voice condition introduced a playful and personalized layer to the experience, though some technical and contextual challenges were pointed out.

4.2.3. Customized voice

Participants' overall feedback on the *customized* Voice was the most positive, which aligns with the quantitative results. Also, it received the highest average rating ($M = 6.19, SD = 1.08$). 76.2% of the participants (16/21) felt that the *customized* voice matched the artifact's appearance and enhanced their engagement. P2 and P6 noted that using a voice that better fits the artifact's appearance makes the task more vivid — *"almost like having a real conversation with the artifact"*. Additionally, some participants (5/21) found the system particularly interesting to use, highlighting the hedonic quality of the system. Two participants (P2 and P10) also mentioned that the voice customization process helped them better retain information about the artifact, suggesting its potential in enhancing memory and connection. However, since the customized interface relied on unlabeled latent axes, users likely incurred additional cognitive effort.

4.2.4. Other suggestions for interaction

For the voice customization tool, most participants emphasized the need for clearer feedback on the five adjustable dimensions (18/21), suggesting that displaying meaningful and descriptive phrases beside the slider would improve their understanding and control of the voice. Other suggestions included incorporating a pause button for voice playback (1/21), improving the speed of voice generation (3/21), and optimizing the interaction interface for better usability (1/21).

5. Discussions

5.1. User preferences and design implications

The primary goal of this study was to explore what types of sounds visitors expect cultural artifacts to produce in virtual museum settings (RQ1) and how different levels of vocal customization affect visitor experience (RQ2). Our findings provide important insights into user preferences and experience design for voice-enabled cultural artifacts. The results indicate a strong preference for the *customized* voice condition (RQ1). Visitors appreciated the opportunity to customize the voice of the artifact and expressed enjoyment in adjusting vocal parameters to suit their own taste or expectations. This aligns with previous findings on user agency and customization in digital environments, suggesting that allowing visitors to co-create or influence the presentation of cultural content fosters a sense of engagement. Participants found the *customized* voice more enjoyable and meaningful, as it gave them a sense of expressive control over the interaction. Specifically, our results showed a significant effect of vocal customization on both pragmatic and hedonic qualities. Results revealed that the *customized* voice condition significantly outperformed both the *default* and *familiar* voice in various dimensions, such as usability, emotional connection, and overall user experience. While the *familiar* voice was moderately preferred over the *default* voice, the differences were not statistically significant. These findings suggest that simply mimicking a familiar person's voice may not be as impactful as giving users the freedom to define the voice themselves.

These findings contribute to broader discussions on personalization and user agency in human–computer interaction. Prior work on customization in digital environments has shown that opportunities for users to shape system outputs increase engagement and perceived relevance [34]. Our results extend this line of research into cultural heritage contexts, showing that adjustable artifact voices are not only technically feasible but also valued by users for enhancing both pragmatic qualities (e.g., usability, clarity) and hedonic qualities (e.g., enjoyment, expressiveness). An important implication is that familiarity does not produce the same benefits as interactive customization. This suggests that agency and expressive control are more important than passive familiarity in shaping user experiences. This resonates with research on co-creation and participatory design in museums [35], where visitors' ability to actively shape their encounters fosters deeper engagement and emotional resonance. From the perspective of cultural heritage communication, our findings highlight the potential of artifact voices as a new modality for storytelling. Giving visitors control over how an artifact "speaks" may increase feelings of presence and connection, resonating with studies on voice embodiment and social presence in VR [36]. This suggests that artifact voices could evolve beyond functional narration to become an expressive medium that reflects visitor interpretation and cultural imagination. Practically, these insights suggest design directions for future museum systems. Voice interfaces for artifacts should not rely solely on generic default voices, nor assume that familiarity with one's own or others' voices guarantees a stronger connection. Instead, interfaces should enable flexible customization through parameters, presets, or AI-driven personalization, so that diverse visitor groups can adapt voices according to their expectations and cultural references. Such flexibility is particularly important in multicultural museum settings, where personalization may support inclusivity and accessibility.

5.2. Limitations and future work

This study offers early insights into user preferences for artifact voice design in virtual museums. However, it also has certain limitations. First, the custom voice tool lacks clearly defined semantic labels for its five adjustable dimensions, which may make it difficult for general users to understand or effectively manipulate the sound parameters. Moreover, the current interface design still require refinement to improve intuitiveness and ease of use, particularly for visitors unfamiliar with audio parameter manipulation. In addition, our participants consist exclusively of university students (aged 19–26) with relatively homogeneous educational backgrounds, which limits the generalizability of our findings to broader and more diverse audiences, such as older adults or museum visitors with different cultural backgrounds. While this homogeneous group was appropriate for an initial exploratory study, future work should validate these findings across more diverse demographic groups to better capture the needs of actual museum visitors. Another limitation concerns the nature of the data collected. The present study relied primarily on self-reported user preferences, which are valuable for capturing perceptions but do not always align with actual behavior in interaction. Future work should therefore complement preference measures with additional data sources such as behavioral logs, interaction traces, and usage analytics. Moreover, our cross-sectional design provides only a snapshot of user attitudes at one point in time. Longitudinal studies will be necessary to examine how preferences evolve as visitors become more familiar with artifact voices and as speech synthesis technologies continue to advance. Future iterations of this research will also benefit from a stronger integration with prior studies in HCI and museum contexts, allowing us to better contextualize the role of artifact voices within ongoing debates

on embodiment, social presence, and participatory design. In the future, we will adopt a human-in-the-loop approach to refine the custom voice tool by engaging a broader group of users with diverse backgrounds in labeling and validating the five sound dimensions, allowing us to derive intuitive, consensus-based descriptors. We also aim to integrate large language models (LLMs) with our tool to support more natural, flexible, and interactive conversations with artifact chatbots.

6. Conclusion

This project represents an initial step toward giving cultural artifacts a voice in virtual museum environments. We explored what types of voices visitors prefer artifacts to have by designing a VR museum that presents 3D models of artifacts accompanied by three voice types: *default*, *familiar*, and *customized*. The *familiar* and *customized* voices were generated using state-of-the-art generative text-to-speech (TTS) models. A user study with 21 participants was conducted to evaluate these voice types. Results indicate a clear preference for *customized* voices, followed by *familiar* voices, with system-generated *default* voices rated the least engaging. These findings highlight the potential of voice customization to enhance engagement and user experience in digital heritage applications. Our study contributes to the ongoing discourse in the field of virtual reality and digital heritage, suggesting that future developments should focus on refining voice customization techniques and exploring additional voice attributes that may affect storytelling capabilities in virtual environments.

CRediT authorship contribution statement

Bingqing Chen: Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Validation. **Wenqi Chu:** Software, Investigation. **Xubo Yang:** Writing – review & editing, Supervision. **Yue Li:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

We thank our participants for their time and efforts. This work is supported by the National Natural Science Foundation of China (62207022).

Data availability

Data will be made available on request.

References

[1] Bem MJ. Effects of sounds on the visitors' experience in museums [Master's thesis], Rensselaer Polytechnic Institute; 2023.
[2] Luo D, Doucé L, Nys K. Multisensory museum experience: an integrative view and future research directions. Mus Manag Curatorship 2024;1–28.
[3] Bertens L, Polak S. Using museum audio guides in the construction of prosthetic memory. J Conserv Mus Stud 2019;17.
[4] J.Bem M, Chabot SR, Brooks V, Braasch J. Enhancing museum experiences: Using immersive environments to evaluate soundscape preferences. J Acoust Soc Am 2025;157:1097–108.
[5] Jakubowski RD. Museum soundscapes and their impact on visitor outcomes [Ph.D. thesis], Colorado State University; 2011.
[6] Chen B, Wen R, Tan S, Li Y. Exploring user preferences for museum guides: The role of chatbots in shaping interactive experiences. In: Proceedings of the extended abstracts of the CHI conference on human factors in computing systems. 2025, p. 1–8.
[7] Garcia IL, Schott E, Gohsen M, Bernhard V, Stein B, Froehlich B. Speaking with objects: Conversational agents' embodiment in virtual museums. In: 2024 IEEE international symposium on mixed and augmented reality. Bellevue, WA, USA: IEEE Computer Society; 2024, p. 279–88. http://dx.doi.org/10.1109/ISMAR62088.2024.00042.
[8] Noh YG, Hong JH. Designing reenacted chatbots to enhance museum experience. Appl Sci 2021;11:7420.
[9] Barbara J, Haahr M. The role of voice in virtual reality interactive narratives. J Interact Narrat 2024.
[10] Wadley G, Carter M, Gibbs M. Voice in virtual worlds: The design, use, and influence of voice chat in online play. Human–Comput Interact 2015;30:336–65.
[11] Abdulrahman A, Richards D. Is natural necessary? human voice versus synthetic voice for intelligent virtual agents. Multimodal Technol Interact 2022;6:51.
[12] Austerjost J, Porr M, Riedel N, Geier D, Becker T, Scheper T, Marquard D, Lindner P, Beutel S. Introducing a virtual assistant to the lab: A voice user interface for the intuitive control of laboratory instruments. SLAS Technol: Transl Life Sci Innov 2018;23:476–82.
[13] Zhang J, Brandstätter K, Steed A. Supporting co-presence in populated virtual environments by actor takeover of animated characters. In: 2023 IEEE international symposium on mixed and augmented reality. IEEE; 2023, p. 940–9.
[14] van Rijn P, Mertes S, Janowski K, Weitz K, Jacoby N, André E. Giving robots a voice: Human-in-the-loop voice creation and open-ended labeling. In: Proceedings of the 2024 CHI conference on human factors in computing systems. 2024, p. 1–34.
[15] Sylaiou S, Fidas C. Virtual humans in museums and cultural heritage sites. Appl Sci 2022;12:9913.
[16] Geigel J, Shitut KS, Decker J, Doherty A, Jacobs G. The digital docent: Xr storytelling for a living history museum. In: Proceedings of the 26th ACM symposium on virtual reality software and technology. 2020, p. 1–3.
[17] Kopp S, Gesellensetter L, Krämer NC, Wachsmuth I. A conversational agent as museum guide–design and evaluation of a real-world application. In: International workshop on intelligent virtual agents. Springer; 2005, p. 329–43.
[18] Duguleană M, Briciu VA, Duduman IA, Machidon OM. A virtual assistant for natural interactions in museums. Sustainability 2020;12:6958.
[19] Al-Taie I, Di Giuseppantonio Di Franco P, Tymkiw M, Williams D, Daly I. Sonic enhancement of virtual exhibits. PLoS One 2022;17:e0269370.
[20] Salo K, Bauters M, Mikkonen T. User generated soundscapes activating museum visitors. In: Proceedings of the symposium on applied computing. 2017, p. 220–7.
[21] Qin HX, Jin S, Gao Z, Fan M, Hui P. Charactermeet: Supporting creative writers' entire story character construction processes through conversation with llm-powered chatbot avatars. In: Proceedings of the 2024 CHI conference on human factors in computing systems. New York, NY, USA: Association for Computing Machinery; 2024, http://dx.doi.org/10.1145/3613904.3642105.
[22] Byrne S. Voicing the museum artefact. J Conserv Mus Stud 2012;10:23–34.
[23] Martin F Alonso, Malfaz M, Castro-González Á, Castillo JC, Salichs MÁ. Four-features evaluation of text to speech systems for three social robots. Electronics 2020;9:267.
[24] Barrault L, Chung YA, Meglioli MC, Dale D, Dong N, Duquenne PA, Elsahar H, Gong H, Heffernan K, Hoffman J, et al. Seamlessm4t: Massively multilingual & multimodal machine translation. 2023, arXiv preprint arXiv:2308.11596.
[25] Tan X, Qin T, Soong F, Liu TY. A survey on neural speech synthesis. 2021, arXiv preprint arXiv:2106.15561.
[26] Kim J, Kong J, Son J. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In: International conference on machine learning. PMLR; 2021, p. 5530–40.
[27] Trichopoulos G, Konstantakis M, Caridakis G, Katifori A, Koukouli M. Crafting a museum guide using chatgpt4. Big Data Cogn Comput 2023;7:148.
[28] Rayward WB, Twidale MB. From docent to cyberdocent: education and guidancein the virtual museum. Arch Mus Inform 1999;13:23–53.
[29] Chen Y, Lyu X, Li T, Gao Z. Li bai the youth: An llm-powered virtual agent for children's chinese poetry education. In: SIGGRAPH Asia 2024 posters. 2024, p. 1–2.
[30] Liang J, Zeng G, Li Y, Dong Y. Artimetravel: Understanding spatial changes in heritage sites over time through web-based augmented reality serious games. In: Proceedings of the extended abstracts of the CHI conference on human factors in computing systems. 2025, p. 1–8.
[31] Trovato G, Trevejo FP, Tordoya AP, Miranda LG, Polo LR. Santo in exhibition–a sacred robot in the profane. In: 2023 32nd IEEE international conference on robot and human interactive communication. IEEE; 2023, p. 1991–6.
[32] Othman MK, Petrie H, Power C. Engaging visitors in museums with technology: Scales for the measurement of visitor and multimedia guide experience. In: Human-computer interaction – INTERACT 2011, vol. 6949, Berlin, Heidelberg: Springer Berlin Heidelberg; 2011, p. 92–9. http://dx.doi.org/10.1007/978-3-642-23768-3_8.
[33] Schrepp M, Hinderks A, et al. Design and evaluation of a short version of the user experience questionnaire (ueq-s). 2017.
[34] Fischer G. Understanding, fostering, and supporting cultures of participation. Interactions 2011;18:42–53. http://dx.doi.org/10.1145/1962438.1962450.
[35] Simon N. The participatory museum. Museum 2.0. 2010.
[36] Biocca F, Harms C, Burgoon JK. Toward a more robust theory and measure of social presence: Review and suggested criteria. Presence: Teleoperators Virtual Environ 2003;12:456–80.