A Character-Based Korean Tourist Dialogue System with Proactive Recommendations and Live2D Embodiment

Jae Young Suh
Republic of Korea
tjwodud04@gmail.com

Mingyu Jeon
Republic of Korea
jkmcoma7@gmail.com

Abstract

This paper presents a character-based Korean tourist assistant prototype that integrates speech recognition, dialogue management, and multimodal embodiment. The system combines client-side voice capture, server-side transcription, retrieval from the Korea Tourism Organization TourAPI, and lightweight policy mechanisms for response generation. Outputs are delivered through a Live2D avatar that provides spoken responses along with recommendation cards and itinerary suggestions. Beyond technical integration, the prototype demonstrates how domain-specific spoken dialogue systems can leverage open APIs and multimodal presentation to offer practical and engaging support in tourism contexts, with implications for user-centered dialogue design.

1 Introduction

Spoken dialogue systems (SDS) have become increasingly sophisticated, yet designing agents that deliver specialized information while fostering natural and engaging interactions remains a core challenge (Jurafsky and Martin, 2025). The tourism domain provides a particularly compelling testbed for this challenge, as tourists often require more than factual retrieval. They seek dynamic recommendations, personalized itineraries, and context-aware guidance that adapt to evolving needs (Prasanna et al., 2025).

Recent studies highlight a shift from reactive exchanges to proactive conversational strategies that anticipate user goals, provide timely suggestions, and embed contextual cues into dialogue flow (Deng et al., 2025; Lu et al., 2025). In tourism, such proactive behaviors are essential for addressing information needs shaped by temporal, spatial, and environmental factors such as time, location, and weather.

This paper presents a prototype of a character-based spoken dialogue system designed to serve as a Korean tourism guide. The system implements a fully integrated, web-based architecture that combines client-side speech capture with a server-side dialogue pipeline. The pipeline retrieves reliable content from the Korea Tourism Organization's TourAPI and applies lightweight proactive policies to generate recommendations. Responses are delivered through a Live2D avatar that vocalizes information while presenting interactive recommendation cards and potential travel routes, thereby creating a multimodal user experience.

The contributions of this work are threefold. First, we demonstrate the practical implementation of a modular SDS pipeline tailored to the Korean tourism domain, integrating a national open API with a character-driven front end. Second, we illustrate how proactive dialogue behaviors, when combined with visual embodiment, enhance both the perceived utility and engagement of a domain-specific conversational system. Third, by situating our prototype within the Korean language and cultural context, we contribute to the development of culturally adaptive dialogue systems that remain lightweight enough for real-time web deployment and live demonstrations.

2 Related Work

2.1 Conversational AI for Tourism and Proactive Dialogue

Conversational AI has become a transformative force in the tourism industry, evolving from simple transactional chatbots to sophisticated, context-aware travel assistants (Alhashmi et al., 2025). Current research emphasizes complex tasks such as dynamic itinerary planning from multi-source data and delivering personalized, culturally aware recommendations (Hu et al., 2020). AI-driven services increasingly compose intricate travel plans that adapt to user preferences, underscoring the value of specialized assistants that can leverage domain-

specific knowledge bases such as national tourism APIs (Pandit et al., 2025). Our work follows this direction by employing the Korea Tourism Organization's TourAPI to provide grounded and reliable information.

A continuing challenge is moving beyond purely reactive question-answering. Proactive dialogue systems seek to address this by anticipating user needs and steering conversations toward successful outcomes (Deng et al., 2023). While many studies explore complex algorithmic approaches, practical implementations can rely on simple but effective policies. Our system illustrates this with a rule-based proactive strategy: once the agent fulfills an initial request for recommendations, it anticipates the next logical step and offers to generate a complete tour course.

2.2 Embodied Agents and Multimodal Interaction

The effectiveness of a conversational agent is strongly shaped by its presentation. Embodied Conversational Agents (ECAs)—systems featuring a visual representation such as an avatar—have consistently been shown to increase user trust, social presence, and engagement compared to text-only or voice-only interfaces (Cassell, 2001). The sense of interacting with a character rather than an abstract system fosters stronger user connection, an effect often linked to the agent's perceived anthropomorphism (Nawaz et al., 2024).

To maximize this effect, modern ECAs employ multimodality by combining verbal output with non-verbal cues. For example, synchronizing an avatar's facial expressions and lip movements with synthesized speech enhances realism and communicative clarity (Chang et al., 2023; Aneja et al., 2021). Our prototype builds on these findings by using a Live2D avatar to provide an expressive presence, combined with a graphical interface that displays interactive recommendation cards. This multimodal approach ensures that information is conveyed in an engaging and easily interpretable format, suitable for real-time, web-based demonstrations.

3 System Overview

Our prototype is a modular, web-based spoken dialogue system for Korean tourism guidance. It integrates a client-side interface for user interaction with a server-side pipeline responsible for dialogue processing, information retrieval, and multimodal response generation. The overall architecture is shown in Figure 1.

3.1 System Workflow

The workflow unfolds in two stages. In Stage 1, the user's voice is captured in the browser using the Web Audio API and sent to a Flask-based server, which transcribes the audio into Korean text. Instead of traditional intent classification, the system converts the transcription into an embedding vector and performs semantic search to retrieve relevant tourist attractions from the Korea Tourism Organization's public TourAPI.

A multimodal response is then generated. The textual output is synthesized into speech and streamed to the client, driving the Live2D avatar's lip-sync animation. At the same time, structured information is rendered as interactive recommendation cards, while a "Yes/No" prompt is displayed to ask whether the user would like to see a related travel course.

Stage 2 is triggered if the user clicks "Yes." A new server request is issued to query travel courses associated with the initially recommended attractions. This two-stage design enables proactive yet incremental dialogue flow.

3.2 User Interface

The user interface, shown in Figure 2, is designed for engaging and intuitive interaction. It consists of two primary panels. On the left, a Live2D avatar serves as a visually embodied conversational agent, providing persistent social presence. The avatar enhances naturalness by delivering spoken responses with synchronized lip movements and expressions corresponding to the dialogue state (e.g., listening, speaking).

The right panel features a chat-style interface displaying conversation history and multimodal outputs. Structured recommendation cards present information—including image, name, description, and link—in a clear format. To minimize perceived latency, textual responses are streamed token by token. User interaction is initiated with a single "Speak" button, ensuring accessibility and simplicity.

4 Preliminary Evaluation

To assess the prototype's viability for live demonstration, we conducted an internal evaluation. The

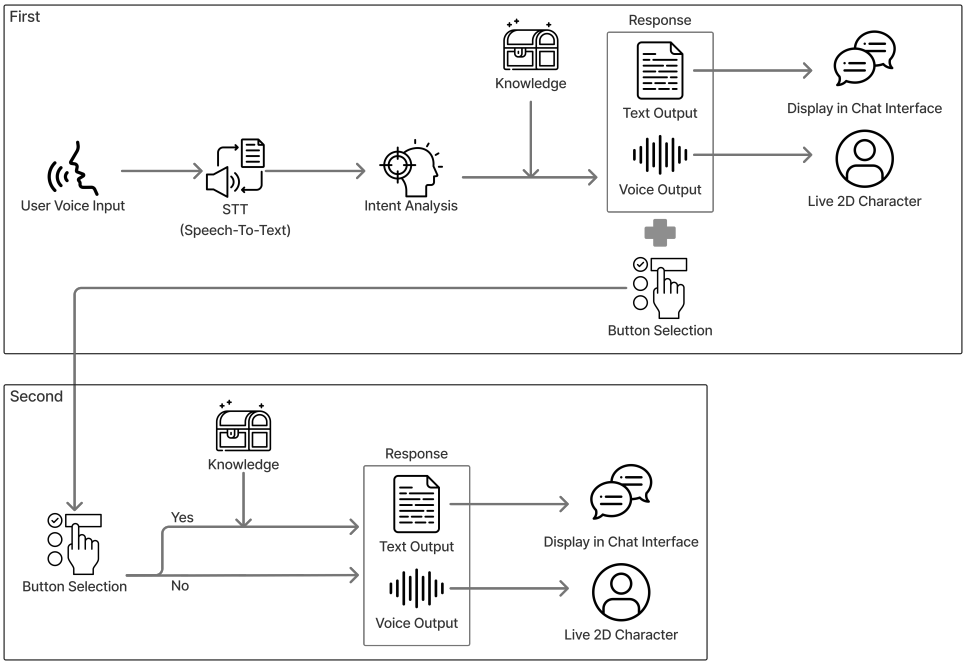

Figure 1: System workflow. Stage 1 (top) processes the initial voice query via semantic search to generate multimodal recommendations. Stage 2 (bottom) is triggered by a button click, leading to a proactive course suggestion based on the initial context.

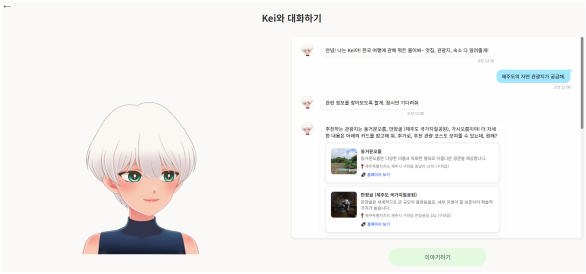

Figure 2: User interface with the Live2D avatar (left) and multimodal chat panel displaying recommendation cards and interactive buttons (right).

test set consisted of 10 spoken queries in Korean for tourist destinations, covering diverse regions and attraction types. The evaluation emphasized the initial interaction, as this stage is critical for the system's ability to interpret user requests and retrieve relevant information.

4.1 Component Performance

The ASR module achieved strong accuracy, with an average Word Error Rate (WER) of 6.4% and Character Error Rate (CER) of 1.3%. Following transcription, the system reliably generated information-rich recommendation cards. Analysis of the 30 cards (3 per query) showed that all (100%) included an image and address, while homepage links were present in 66.7% of cases, reflecting data availability in the source API. Table 1 illustrates a representative QA example from the test set.

Table 1: Example from the evaluation set.

Component	Content / Example
User Query (Spoken)	부산에서 가족과 함께 즐길 수 있는 관광지 추천해줘 (Recommend a tourist spot in Busan for my family.)
System Response (Spoken)	추천하는 관광지는 금곡동 율리 바위그늘유적, 기장군반려동물문화축제, 금정산성 동문이야!... (The recommended spots are Yulli Rock Shelter...)
Generated Card (1 of 3)	**Name:** 금곡동 율리 바위그늘유적 (Geumgok-dong Yulli Rock Shelter) **Address:** 부산광역시 북구 금곡동 산24 (San 24, Geumgok-dong, Buk-gu, Busan) **Image URL:** https://tong.visitkorea.or.kr/cms/... **Homepage:** https://www.heritage.go.kr/heri/cul/...

4.2 System-Level Performance

End-to-end latency, measured from the end of user speech to the start of the avatar's response, averaged 8.36 seconds. To reduce perceived delay, the system displays a textual status message (e.g., "관련 정보를 찾아보도록 할게. 잠시만 기다려줘."(I'll look up the relevant information. Please wait a moment.)) approximately 3 seconds after speech cessation. approximately 3 seconds after speech cessation. The results, summarized in Table 2, indicate that the prototype is reliable and

suitable for real-time demonstration.

Table 2: Summary of Performance Evaluation.

Category	Metric	Value
ASR Performance	Avg. Word Error Rate (WER)	6.4%
	Avg. Character Error Rate (CER)	1.3%
Recommendation Card Content	Image Coverage	100%
	Address Coverage	100%
	Homepage Link Coverage	66.7%
System Latency	Avg. End-to-End Response Time	8.36 sec

These findings suggest that while the prototype demonstrates reliable functionality for live demonstration, future work should focus on reducing response latency and addressing incomplete metadata in the source API to further improve user experience.

5 Discussion

Our preliminary evaluation provides key insights for designing embodied dialogue systems in the tourism domain. The results confirm that a modular architecture integrating speech recognition, API retrieval, and multimodal output is viable for web-based demonstration. The high accuracy of the ASR module (6.4% WER) forms a critical foundation, enabling the system to reliably interpret user intent and activate proactive dialogue policies, which remain central to effective conversational AI (Deng et al., 2023).

From a user experience perspective, the Live2D avatar enhances engagement by fostering a sense of social presence, a well-documented factor in the adoption of conversational agents (Xu et al., 2024). This embodied presence, together with visually rich recommendation cards, produces a multimodal interaction that feels more tangible and culturally adaptive than text-only outputs (Aneja et al., 2021; Alhashmi et al., 2025). The anthropomorphic qualities of the avatar further align with findings that human-like features can positively influence trust and satisfaction in AI companions (Nawaz et al., 2024).

At the same time, the prototype highlights practical challenges. The average end-to-end latency of 8.4 seconds, inherent to the sequential, cloud-based pipeline of STT, semantic search, and TTS services, remains a primary concern. To mitigate this, the system displays a brief status message at the 3-second mark, which functions as a conversational filler and helps manage user expectations (Skantze, 2021). In addition, reliance on an external API couples system performance with the completeness of source data, as reflected in the occasional absence of homepage links. Future work should explore strategies for handling such inconsistencies more gracefully.

In conclusion, this prototype functions as a concrete case study for integrating proactive dialogue with an embodied interface in a specialized domain such as Korean tourism. Key directions for future development include enhancing dialogue management capabilities by implementing dialogue state tracking for contextual follow-up queries (e.g., "그 지역 날씨는 어떤지 알려줄래?"(What is the weather like there?)) and supporting multi-step tasks such as comparative trip planning.

6 Conclusion

This paper presented and evaluated a character-based spoken dialogue system for Korean tourism, offering a concrete blueprint for integrating proactive dialogue strategies, a national open API, and a visually embodied agent within a modular web-based architecture. Our preliminary evaluation confirmed reliable end-to-end performance, with accurate speech recognition and consistent generation of multimodal recommendations. These findings highlight that the integration of proactive behaviors and visual embodiment can enhance both the utility and engagement of domain-specific conversational agents, making the system demonstration-ready.

Despite this promise, key challenges remain, most notably system latency and the absence of advanced multi-turn dialogue management. Future work will focus on optimizing the processing pipeline to reduce response times and extending the agent's conversational capabilities to support more complex interactions. Addressing these challenges will contribute to developing more natural and effective embodied conversational agents for practical applications in real-world settings.

References

Saadat Alhashmi, Mohamed Aboelmaged, Ibrahim Hashem, and Muhammad Bilal. 2025. Two decades of chatbot research in tourism and hospitality: Bibliometric analysis and future directions. *Sustainable Communities*, 2.

Deepali Aneja, Rens Hoegen, Daniel McDuff, and Mary Czerwinski. 2021. Understanding conversational and expressive style in a multimodal embodied conversational agent. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

Justine Cassell. 2001. Embodied conversational agents: representation and intelligence in user interfaces. *AI Mag.*, 22(4):67–83.

Che-Jui Chang, Samuel S Sohn, Sen Zhang, Rajath Jayashankar, Muhammad Usman, and Mubbasir Kapadia. 2023. The importance of multimodal emotion conditioning and affect consistency for embodied conversational agents. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, page 790–801, New York, NY, USA. Association for Computing Machinery.

Yang Deng, Wenqiang Lei, Wai Lam, and Tat-Seng Chua. 2023. A survey on proactive dialogue systems: problems, methods, and prospects. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI '23.

Yang Deng, Lizi Liao, Wenqiang Lei, Grace Hui Yang, Wai Lam, and Tat-Seng Chua. 2025. Proactive conversational ai: A comprehensive survey of advancements and opportunities. *ACM Transactions on Information Systems (TOIS)*, 43(3).

G. Hu, Y. Qin, and J. Shao. 2020. Personalized travel route recommendation from multi-source social media data. *Multimedia Tools and Applications*, 79:33365–33380.

Daniel Jurafsky and James H. Martin. 2025. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models*, 3rd edition. Online manuscript released August 24, 2025.

Jiahao Lu, Wenjie Wang, Wenqiang Chen, and Yang Deng. 2025. Protod: Proactive task-oriented dialogue system based on chain-of-plan-execution. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 8945–8956. Association for Computational Linguistics.

Raja Nawaz, Sajjida Reza, and Bilal Sarwar. 2024. Chatbot anthropomorphism and its influence on trust and satisfaction via perceived value. *Journal of Social Organizational Matters*, 3:543–568.

Ameet Pandit, Philip Rosenberger III, Shah Miah, Violetta Wilk, and Abdulaziz Alharbi. 2025. Artificial intelligence-enabled conversational agents in tourism hospitality: a systematic literature review future research directions. *Asia Pacific Journal of Tourism Research*.

Akshara Prasanna, P. Pushparaj, and Bijay Prasad Kushwaha. 2025. Conversational ai in tourism: A systematic literature review using tcm and ado framework. *Journal of Hospitality and Tourism Management*, 64:101310.

Gabriel Skantze. 2021. Turn-taking in conversational systems and human-robot interaction: A review. *Computer Speech Language*, 67:101178.

Han Xu, Rob Law, Jon Lovett, Jian Ming Luo, and Lu Liu. 2024. Tourist acceptance of chatgpt in travel services: the mediating role of parasocial interaction. *Journal of Travel & Tourism Marketing*, 41(7):955–972.