Key Challenges in Multimodal Task-Oriented Dialogue Systems: Insights from a Large Competition-Based Dataset

Shiki Sato[1], Shinji Iwata[1], Asahi Hentona[1], Yuta Sasaki[2], Takato Yamazaki[3], Shoji Moriya[4], Masaya Ohagi[3], Hirofumi Kikuchi[5], Jie Yang[5], Zhiyang Qi[6], Takashi Kodama[7], Akinobu Lee[8], Masato Komuro[9], Hiroyuki Nishikawa[10], Ryosaku Makino[5], Takashi Minato[11,12], Kurima Sakai[11], Tomo Funayama[11], Kotaro Funakoshi[2], Mayumi Usami[13], Michimasa Inaba[6], Tetsuro Takahashi[14], Ryuichiro Higashinaka[15]

[1]CyberAgent, [2]Institute of Science Tokyo, [3]SB Intuitions Corp., [4]Tohoku University, [5]Waseda University, [6]The University of Electro-Communications, [7]NII LLMC, [8]Nagoya Institute of Technology, [9]Chiba University, [10]Meikai University, [11]ATR, [12]RIKEN, [13]Tokyo University of Foreign Studies, [14]Kagoshima University, [15]Nagoya University

Correspondence: sato_shiki@cyberagent.co.jp

Abstract

Challenges in multimodal task-oriented dialogue between humans and systems, particularly those involving audio and visual interactions, have not been sufficiently explored or shared, forcing researchers to define improvement directions individually without a clearly shared roadmap. To address these challenges, we organized a competition for multimodal task-oriented dialogue systems and constructed a large competition-based dataset of 1,865 minutes of Japanese task-oriented dialogues. This dataset includes audio and visual interactions between diverse systems and human participants. After analyzing system behaviors identified as problematic by the human participants in questionnaire surveys and notable methods employed by the participating teams, we identified key challenges in multimodal task-oriented dialogue systems and discussed potential directions for overcoming these challenges.

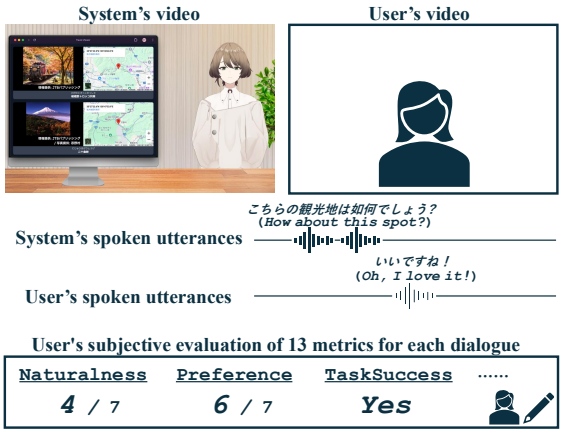

Figure 1: Example of a dialogue in our competition-based dataset. It contains 1,865 minutes of Japanese audio-visual task-oriented dialogues and subjective dialogue evaluation results. Maps Data: Google.

1 Introduction

Task-oriented dialogue systems are in high demand in both academia and industry (Qin et al., 2023; Ni et al., 2023). In recent years, particular attention has been given to multimodal task-oriented dialogue systems that enable more natural and richer real-time interactions through spoken utterances and visual representations of both participants (i.e., real-time camera input capturing the user, as well as output provided via system avatar video or embodied robot actions) (Valizadeh and Parde, 2022; Chen et al., 2025). In this paper, we refer to these dialogues utilizing both audio and visual modalities as "audio-visual dialogues."

Despite significant attention in this area, the challenges involved in audio-visual task-oriented dialogue between humans and systems have not been sufficiently explored or shared within the research community. Consequently, researchers are forced to define improvement directions individually without a clearly shared roadmap. This situation is mainly due to the lack of datasets for audio-visual task-oriented dialogues between humans and systems. Without detailed analyses of such datasets, it is difficult to precisely identify the key factors influencing user experience and task performance, particularly given the complex interactions across multiple modalities.

As an initial step toward promoting a shared understanding of key challenges in audio-visual task-oriented dialogues, we organized a competition at the 7th Dialogue System Live Competition (DSLC7) (Sato et al., 2025) in which we constructed a dataset of Japanese audio-visual task-

Proceedings of the 26th Annual Meeting of SIGDIAL, pages 449–460
Aug 25–27, 2025.

oriented dialogues between various participating systems and humans, as shown in Fig. 1. This competition consisted of two stages: a preliminary round and a final round. In the preliminary round, 94 human evaluators interacted with the participating systems and provided subjective evaluations after each dialogue. As a result, we collected 1,865 minutes of publicly shareable data comprising 257 audio-visual task-oriented dialogues along with their corresponding subjective evaluation results. In the final round, the top three systems selected from the preliminary evaluations engaged in a total of six dialogues with the representatives we invited. The six dialogues were observed and evaluated by an audience of approximately 110 people, including dialogue researchers. Additionally, we surveyed the audience before and after the final round, enabling us to analyze changes in their perceptions regarding key challenges in audio-visual dialogue systems throughout the competition.

In this paper, we first describe the competition we organized and the dataset created during the competition. We then analyze the competition-based dataset to identify key challenges in multimodal task-oriented dialogue systems involving audio and visual modalities. Finally, we discuss potential approaches to address these challenges based on notable methods implemented by participating teams. Our dataset and the codebase of system templates provided to participating teams (Section 3.1) will be accessible from the DSLC7 project page.[1]

2 Related Work

Multimodal task-oriented dialogue datasets. Various multimodal task-oriented dialogue datasets have been developed in response to the growing interest in multimodal dialogue. However, these datasets generally consist of dialogues that lack audio (i.e., spoken utterances) or visual modalities (i.e., videos of dialogue participants) (Walker et al., 2001; Raux and Eskenazi, 2004; Williams et al., 2013; Henderson et al., 2014; Saha et al., 2018; Thomason et al., 2020) or that have been collected either from human-human interactions or via the Wizard-of-Oz approach (Hemphill et al., 1990; Thompson et al., 1993; Horiuchi et al., 1999; Spanger et al., 2012; Jayagopi et al., 2013; Gorisch et al., 2014; Kim et al., 2016; Hayakawa et al., 2016; Byrne et al., 2019; Canévet et al., 2020; Okamoto et al., 2020; Kim et al., 2021; Komatani and Okada, 2021; Kim et al., 2022; Eijk et al., 2022; Inaba et al., 2024; Zhang et al., 2023; Soltau et al., 2023; Si et al., 2023). Consequently, the analysis of audio-visual multimodal task-oriented dialogues between actual systems and human users remains limited. In this study, we constructed an audio-visual task-oriented dialogue dataset containing interactions between diverse systems and diverse human users through organizing a competition.

Dialogue system competitions. Organizing competitions for dialogue systems is vital, not only for driving technological advancements but also for clarifying and sharing key research challenges. To date, several competitions focusing on dialogue systems have been conducted, such as the Dialog System Technology Challenges (Soltau et al., 2023),[2] Alexa Prize,[3] and Dialogue System Live Competitions (Higashinaka et al., 2024). However, competitions that compare the overall performance of audio-visual task-oriented dialogue systems based on human evaluation remain scarce. A notable exception is the Dialogue Robot Competition, where audio-visual task-oriented dialogues between robots and humans have been conducted (Minato et al., 2023). However, due to the nature of this competition—specifically, evaluating the performance of systems deployed in actual retail environments—the dialogue dataset is not made publicly available and has not been sufficiently analyzed. In this study, we analyze our publicly available competition-based dataset to uncover challenges in audio-visual task-oriented dialogue systems.

3 Organization of Competition

As a first step towards identifying and sharing key challenges in audio-visual task-based dialogue systems, we organized a competition at DSLC7 in Japan. Through this competition, we collected Japanese dialogues between participating systems and humans and associated evaluation data. Constructing a dataset through a competition enabled us to collect dialogues of diverse systems.

The competition consisted of preliminary and final rounds. In the preliminary round, dialogue systems from participating teams engaged in face-to-face interactions with human evaluators. After

[1] https://sites.google.com/view/dslc7.

[2] https://dstc11.dstc.community/past-challenges.

[3] https://www.amazon.science/alexa-prize

each dialogue, human evaluators provided subjective evaluations of the systems. Systems that received high subjective evaluation scores advanced to the final round. In the final round, these selected dialogue systems interacted with dialogue interlocutors designated by us in front of an audience that included dialogue system researchers. The audience evaluated the systems based on the demonstrated dialogues as a third party.

3.1 Task Settings

A key objective of this competition is to collect diverse interactions that comprehensively reveal challenges in audio-visual task-oriented dialogue systems. To achieve this, it is crucial to involve human interlocutors with a broad range of attributes. However, simply recruiting diverse participants is not sufficient; for participants to naturally engage in dialogue, the task setting itself must be designed in a way that encourages natural conversation regardless of individual background or experience. With this in mind, we designed a "Tourist Spot Selection Task," inspired by the Dialogue Robot Competition. As Inaba et al. (2024) mentioned, travel planning is considered a dialogue topic that people with diverse attributes can naturally participate in.

Tourist spot selection task. In this task, a dialogue system acts as a travel agency's counter salesperson, suggesting tourist spots to the human interlocutor. Through a dialogue, the system aims to determine a primary tourist spot in Japan that aligns with two travel objectives specified by the interlocutor. A screenshot of a dialogue system is shown in the upper left of Fig. 1. The system is allowed to display images and maps (via Google Maps[4]) of up to four tourist spots to the interlocutor through a virtual monitor (hereafter "Travel Viewer").

Travel objectives. The interlocutor selects two travel objectives before the dialogue. The selection process is conducted in two stages. First, the interlocutor randomly draws five "Objective Cards," each describing different travel objectives, from a set of 52 unique cards we prepared for this competition. This initial random draw ensures diversity among the objectives that appear in the dataset, reflecting a wide range of travel motivations. Then, from these five cards, the interlocutor selects two objectives that they feel would be natural and easy to discuss as their own travel goals. This second step allows interlocutors to choose objectives that are personally relatable, thereby facilitating more natural and engaging dialogue. Examples of these objectives include "Taking attractive photographs at scenic spots" and "Organizing a company trip." The travel objectives remain undisclosed to participating teams, requiring them to develop systems that are flexible enough to accommodate any given objective.

Inputs and outputs. The systems submitted to this competition can use the human interlocutor's voice and frontal video as input. The output is the specified CG avatar's motion commands and its synthesized voice. The competition does not restrict the manner of turn-taking, and the system and the interlocutor are allowed to interrupt each other's speech.

System requirements. The full specifications of all requirements are provided in Appendix A. Requirements include, for example, designated CG avatars, CG avatar software, and a text-to-speech API.

Evaluation metrics. Table 1 summarizes the evaluation metrics used in the evaluation. Based on these metrics, each system's performance is ultimately represented by two aggregate scores: the Satisfaction score and the Completion score. The Satisfaction score is obtained by averaging the ten metrics that capture the dialogue-experience perspective for each evaluator, as listed in the upper half of Table 1, and then averaging these values across all evaluators. The Completion score is calculated as the product of (i) the proportion of evaluators who answered "Yes" to Task Success (Suc) and (ii) the mean of their objective-achievement ratings, computed by first averaging of each evaluator's Obj1 and Obj2 scores (which assess, for each Objective Card, whether the determined tourist spot met the stated purpose; see Table 1) and then averaging these means across the evaluators who answered "Yes" to Suc. All metrics in Table 1, except those requiring binary answers ("Yes" or "No"), are rated on a 7-point Likert scale. These metrics follow those previously employed in the Dialogue Robot Competition.

System template. As an example system for the submission to our competition, we provided participating teams with a system that meets the specified

[4] https://developers.google.com/maps.

Perspective	Evaluation metric	Abbreviation
Satisfaction	Was the system's tourist information sufficient? (Information Sufficiency)	Inf
	Was the dialogue with the system natural? (Naturalness)	Nat
	Was the system behavior appropriate? (Appropriateness)	App
	Was the system behavior preferred? (Preference)	Ple
	Was the dialogue with the system satisfactory? (Satisfaction)	Sat
	Did you find the system trustworthy? (Trustworthiness)	Tru
	Was the system's information useful in selecting a tourist spot? (Usefulness)	Use
	Was the system's information reliable? (Reliability)	Rel
	Would you like to visit this travel agency again? (Return Intention)	Ret
	Did you get excited about traveling? (Expectation for Travel)	Exp
Completion	Did you decide on a tourist spot? [Yes/No] (Task Success)	Suc
	Did the tourist spot meet the purpose of the first Objective Card?	Obj1
	Did the tourist spot meet the purpose of the second Objective Card?	Obj2

Table 1: List of evaluation metrics. Except for "Task Success," which is answered as "Yes" or "No," all metrics are evaluated on a 7-point scale.

requirements based on Remdis (Chiba et al., 2024), a platform designed for developing real-time multimodal dialogue systems. Participants were not required to base their development on this template as long as they satisfied the aforementioned requirements.

3.2 Preliminary Round

Participating teams. The competition initially received entries from 12 teams. Of these, two teams withdrew before the preliminary round's evaluation. Additionally, our screening process determined that two more teams could not be evaluated, leaving eight teams qualified for the preliminary round. Finally, nine systems participated in the preliminary round's evaluation, with the addition of our baseline system.

Procedures of the preliminary round. The human evaluators first received an explanation of how to interact with and assess each dialogue system. Evaluators who consented to participate after receiving the explanation interacted with the assigned participating systems and evaluated each system immediately following each interaction. Each evaluator repeated this task—dialogue followed by evaluation—up to four times, thus evaluating up to four distinct systems. To eliminate potential effects arising from the order in which dialogues were conducted, we employed a counterbalancing design. Each dialogue was carried out in a private room, with the evaluator interacting face-to-face with the system (see Fig. 2).

Comment-based feedback. In addition to the rating-based evaluation for the metrics introduced in Section 3.1, evaluators were asked to provide free comments highlighting positive and negative aspects of the system for each dialogue as supplementary information.

Figure 2: Scene from the preliminary round.

Collection of dialogues. With the consent of evaluators and participating teams, we collected records of the dialogues conducted during the preliminary round. These consisted of audio and video recordings from both the systems and the human evaluators, in addition to system logs.

3.3 Final Round

In the final round, we demonstrated task-oriented dialogues between the representatives we invited and the three top-rated systems from the preliminary round[5] in front of an audience of approximately 110 people, including many dialogue researchers. The audience evaluated each system's dialogue performance as a third party, which determined the final rankings of the top three teams in the competition. For details of the dialogue demonstration, please refer to Appendix B.

The final round had two additional objectives

[5]We selected three systems that scored particularly high in both Satisfaction and Completion in the preliminary round.

beyond determining the rankings of the competing systems. The first was to enable the audience to observe and understand the current capabilities and challenges of audio-visual task-oriented dialogue systems. The second objective was to widely share the challenges recognized by the audience with the research community. With this in mind, both before and after the final round, we surveyed the audience about the recognition of challenges in audio-visual dialogue systems (hereafter, final round fixed-point questionnaire).[6] The questionnaire requested participants to select three challenges they considered most significant from our list (Table 3). In conducting the same questionnaire before and after the event, we aimed to clarify how perceptions of the challenges changed due to the competition.

4 Collected Data

Preliminary round data. We collected 257 dialogues between nine systems and 94 human evaluators, totaling 1,865 minutes, along with subjective evaluation results provided by evaluators. Each dialogue consists of a quadruple: a frontal video of the interlocutor, a video of the dialogue system avatar and the Travel viewer, separated audio recordings of both dialogue speakers, and system input/output logs. To the best of our knowledge, this is the first large-scale dataset of audio-visual task-oriented dialogues between humans and systems. More detailed information on the dataset is provided in Appendix C.

Final round data. We collected six dialogues—two dialogues each from three selected systems—and obtained objective evaluation results from the audience members for each dialogue. The components of each dialogue are identical to those of the preliminary data. In addition, we included the results of the final round fixed-point questionnaire from 86 members of the final round audience.

5 Analysis

We collected a large-scale dataset comprising audio-visual task-oriented dialogues between diverse dialogue systems and human evaluators and corresponding evaluation results. By analyzing them, we investigate the current key challenges in audio-visual task-oriented dialogues.

Numerous analytical perspectives are possible in audio-visual task-oriented dialogues, where multiple modalities intricately interact to influence user experience and task accomplishment. In this paper, we focus on the issues that were frequently pointed out in the preliminary and final round questionnaires. These issues may have a substantial impact on the dialogues, as they drew notable attention from evaluators in the preliminary round and from the audience in the final round. Given that our dataset represents the first large-scale collection of audio-visual task-oriented dialogues between humans and systems, our analysis specifically emphasizes the issues associated with the audio-visual modality rather than the semantic content of the dialogues.

5.1 Summary of Questionnaire Results

Free comments in the preliminary round. Among the 257 dialogues conducted in the preliminary round, free-text comments were provided for 165. Among these, 128 dialogues contained comments that described specific problems. The first author read every comment and segmented them into discrete problem statements, with each segment counted as one issue. Applying this manual procedure resulted in a total of 161 issues. These 161 issues were then manually grouped into the categories shown in Table 2. Notably, more than half of these comments referred not to the semantic content of the system's utterances but to system behaviors involving audio or visual aspects, such as response timing or the avatar's facial expressions. These issues can be broadly classified into five categories: "Timing of responses," "Length of responses," "Backchanneling," "Speech recognition / Synthesis," and "Facial expressions of avatar."

Final round fixed-point questionnaire. Table 3 presents the results of the fixed-point questionnaire to the audience during the final round. These results show that, prior to the final round, audience votes were relatively evenly distributed across most of the listed issues, with some minor biases. In contrast, after the final round, there was a notable increase in votes regarding the naturalness of turn-taking, while votes concerning deficiencies in facial expressions, gestures, and system actions during non-speech periods decreased sharply.

[6] Note that this competition was held concurrently with another competition involving audio-visual dialogue systems addressing a different task (Takahashi et al., 2025), and in a strict sense, the questionnaire also included feedback regarding that competition. However, since both competitions were conducted in the same format, except for the dialogue situations and evaluation metrics, the questionnaire results are considered to highlight the challenges identified in our competition.

Category	Issue	Frequency
Timing of responses	Response delay	17
	Interruption of human utterances, Unnatural response timing	10
Length of responses	Too long responses, Rapid-fire responses	14
Backchanneling	Inappropriate frequency or timing of backchanneling	7
Speech recognition / Synthesis	Incorrect speech recognition results	12
	Mispronunciations (especially of kanji characters)	9
Facial expressions of avatar	Rigid or subdued facial expressions	5
Others	Issues with a frequency of less than five	15
	Issues related to dialogue contents (e.g., failure to reflect tourism purposes in the system's recommendation)	72

Table 2: Issues mentioned in the subjective evaluators' free comments during the preliminary round and their frequency. Only issues involving vision or audio are shown in detail.

Issue	BF	AF
Unnatural turn-taking	58	75
User and environment unawareness	54	60
Slow operation, Poor real-time performance	51	52
Insufficient facial expressions and gestures	30	19
Insufficient actions during non-speech	28	19
High processing costs	23	17
Misaligned gestures	13	10
Response with inappropriate acoustics	1	0
Others	0	6

Table 3: Results of the final round fixed-point questionnaire from 86 participants. Respondents selected three issues from the list that they felt were most important. BF and AF represent the number of respondents who selected the issue before and after the final round, respectively.

During the preliminary round, evaluator comments primarily pointed out issues related to response timing, response length, backchanneling, speech recognition/synthesis, and the avatar's facial expressions. Among these, system behavior concerning turn-taking drew particular attention in the final round. In the following sections, we analyze our large-scale dataset collected during the preliminary round to investigate how these behaviors impact user experience and evaluation results. We also discuss potential approaches for addressing them based on notable methods by some participating teams. Detailed statistical values for subsequent analyses are provided in Tables 4 and 5.

5.2 Analysis: Timing of Responses

5.2.1 Response Delay

Among the issues raised during the preliminary round, response delay was the most frequently mentioned. To quantitatively assess its impact, we calculated Spearman's rank correlation coefficients between the cumulative duration of pauses—from the end of human utterances to the beginning of system responses—of each dialogue and evaluators' ratings across all evaluation metrics for the dialogue. The results revealed statistically significant negative correlations between the cumulative pause duration and four evaluation metrics—Nat, App, Ple, and Sat—from the Satisfaction perspective, as well as task completion ratings (Table 4).

These findings provide quantitative evidence that response delays can affect user experience and task completion in audio-visual task-oriented dialogues. This observation is intriguing in light of prior reports suggesting that humans tend to be tolerant of response delays in task-oriented dialogues (Peng et al., 2020).

While the fundamental solution to this issue is clearly to improve the response generation speed, current techniques face inherent limitations in achieving such acceleration. Consequently, some teams implemented strategies such as inserting fillers or thinking gestures during response preparation to mitigate perceived response delays. Previous studies indicate that such fillers and gestures can reduce human-perceived delays (Kum and Lee, 2022). In fact, one evaluator explicitly commented positively, stating, "It was good that the system did not remain silent while it was thinking." However, we observed a case where these strategies yielded a negative impact. In a free comment on one dialogue in the preliminary round, one evaluator explicitly mentioned being bothered by the repeated filler "um" in the same dialogue. Indeed, we found that the system generated 13 instances of "um" and similar fillers during that dialogue, potentially causing annoyance due to their repetitiveness.

Dialogue feature	Inf	Nat	App	Ple	Sat	Tru	Use	Rel	Ret	Exp	Tsk
Total response delay (§5.2.1)	–.058	–.137*	–.151*	–.160*	–.124*	–.087	–.077	–.053	–.077	–.025	–.192*
No. of interruptions (§5.2.2)	–.203*	–.239*	–.250*	–.243*	–.202*	–.175*	–.186*	–.161*	–.148*	–.150*	–.232*
Average response length (§5.3)	–.175*	–.165*	–.220*	–.208*	–.224*	–.191*	–.248*	–.201*	–.215*	–.197*	–.223*
No. of backchannels (§5.4)	–.273*	–.267*	–.280*	–.261*	–.218*	–.169*	–.175*	–.193*	–.194*	–.185*	–.071
No. of backchannels (<15) (§5.4)	.055	.176	.258*	.185	.139	.134	.081	.094	.167	.073	.364*
No. of expression changes (§5.6)	.165*	.182*	.196*	.217*	.220*	.158*	.155*	.204*	.208*	.166*	.150*

Table 4: Spearman's rank correlation coefficients between each dialogue feature and the evaluation scores. Tsk is defined as a single Task Completion metric, used here for correlation analysis, by aggregating the three Completion-related metrics (Suc, Obj1, and Obj2) into one score. For each dialogue, if Task Success is "Yes," Tsk equals the average of Obj1 and Obj2; if "No," Tsk is set to 0. Note that * indicates significance at $p < 0.05$.

Dialogue subset	Inf	Nat	App	Ple	Sat	Tru	Use	Rel	Ret	Exp	Suc	(Obj1+Obj2) / 2	Tsk
w/ SR errors (§ 5.5.1)	4.75	4.00	3.92	4.50	4.25	4.17	5.00	4.75	4.17	4.75	0.92	5.58	5.64
w/o SR errors	5.22	4.67	4.81	4.91	4.70	4.94	5.31	5.30	4.42	5.00	0.94	5.08	5.26
w/ SS errors (§ 5.5.2)	5.67	5.22	5.00	4.89	4.67	4.78	5.44	5.56	4.22	5.11	1.00	4.67	4.67
w/o SS errors	5.18	4.61	4.76	4.89	4.68	4.90	5.28	5.26	4.42	4.99	0.93	5.12	5.30

Table 5: Average user evaluation scores for each dialogue subset. "w/ SR errors" and "w/o SR errors" indicate subsets of dialogues in which speech recognition errors were reported or not, respectively, while "w/ SS errors" and "w/o SS errors" indicate the corresponding subsets for speech synthesis errors.

As tasks increase in complexity, dialogue durations are expected to lengthen, increasing the count of response delays and consequently increasing fillers and gestures. Given that audio-visual dialogues are gradually handling more complex tasks, careful consideration should be given to ensuring diversity in fillers and gestures to mitigate the potential negative impacts of these approaches.

5.2.2 Interruption of Human Utterances, Unnatural Response Timing

While delays in system responses were frequently mentioned, some evaluators also reported interruptions of their utterances by the system and the reverse phenomenon. We calculated Spearman's rank correlation coefficients between the number of interruptions per dialogue (instances where both system audio and human audio occurred) and evaluation scores from the preliminary round for each dialogue. The analysis confirmed that the number of interruptions exhibited statistically significant negative correlations with all evaluation metrics, both from the Satisfaction and Task completion perspectives (Table 4). This result suggests that interruptions, in addition to response delays, represent a critical issue with a serious impact on user satisfaction and task completion.

One team attempted to reduce these interruptions by utilizing visual cues. Specifically, this team implemented a feature in which the system avatar extended its hand toward the interlocutors, indicating the appropriate timing for them to begin speaking. Indeed, the average number of interruptions per dialogue for this team's system was 7.1, which is lower than the 8.4 interruptions observed on average in other teams' systems.

However, even in dialogues with this system, utterance interruptions were not entirely eliminated. One potential reason is that evaluators might not have noticed the avatar's hand movements. In this dialogue task, essential task-related information is visually provided through the Travel Viewer interface. Providing such visual information might reduce evaluators' attention toward the avatar, thereby limiting the effectiveness of visually guided turn-taking strategies.

In dialogues enriched with extensive visual information, simply transplanting turn-taking cues proposed in previous audio-visual studies, such as an avatar extending its hand to signal the floor, may have only limited effectiveness.

5.3 Analysis: Length of Responses

Another frequently mentioned issue was the excessively long duration of uninterrupted system utterances. Upon detailed observation of these comments, it became evident that long utterances were particularly noted during explanations about tourist spots. This suggests that system utterances became notably lengthy when attempting to deliver information about tourist spots in a single utterance.

Indeed, a correlation analysis between the du-

ration of system utterances and human evaluation scores from the preliminary round suggested that extremely long system utterances significantly impact user experience and task performance. Specifically, all evaluation metrics for both user satisfaction and task completion showed statistically significant negative correlations with the average duration (in seconds) per system utterance during interactions (Table 4).

This result is interesting when considered alongside previous studies reporting relatively high human tolerance towards verbose system responses (Whittaker et al., 2003). Although partially overlapping with the discussion in the previous section, one possible factor contributing to the discrepancy from these earlier findings could be the relatively longer dialogue duration involved in this task. Given that prolonged interactions may occasionally lead to decreased interlocutors' concentration, future studies could explore methods to alleviate users' cognitive load, such as breaking down or shortening information delivery during dialogues.

5.4 Analysis: Backchanneling

Several evaluators commented that the systems produced too many backchannels or that their timing appeared unnatural. Indeed, calculating Spearman's rank correlation coefficient between the number of backchannels (the total number of system utterances of fewer than six characters during user utterances) per dialogue and evaluator ratings of each dialogue revealed that this feature exhibited statistically significant negative correlations with all the evaluation metrics from the Satisfaction perspective (Table 4).

Conversely, when analyzing only dialogues with fewer than 15 backchannels, we found a statistically significant positive Spearman's rank correlation between the frequency of backchannels and evaluator ratings of Appropriateness (Table 4). This suggests that appropriately frequent backchanneling can have a beneficial impact on user experience. These results suggest that the tendencies previously observed in non-task-oriented dialogues—where both excessively frequent and insufficient backchannels degrade interaction quality (Poppe et al., 2011)—can also apply to task-oriented dialogues. Thus, achieving natural and desirable task-oriented dialogues may require an optimal frequency of backchannels.

5.5 Analysis: Speech Recognition/Synthesis

Errors in speech recognition and synthesis still occurred, and numerous comments regarding these errors were received in the preliminary round.

5.5.1 Speech Recognition

In particular, dialogues in which speech recognition errors were reported[7] exhibited lower scores across all Satisfaction perspectives' metrics averaged among all evaluators, as well as a lower percentage of evaluators answering "Yes" for Task Success, compared to dialogues without such errors (Table 5). Notably, the evaluation for Appropriateness had an average score of 3.92, which is significantly lower than the 4.82 recorded for dialogues without reported speech recognition errors (one-tailed Welch's t-test, $p = 0.043$). These results suggest that speech recognition errors negatively impact user experience.

Two teams adopted a simple yet potentially effective approach to mitigate task failures arising from speech recognition errors: explicitly asking interlocutors to confirm the accuracy of information captured by the system. Indeed, one evaluator positively commented on one of these teams, highlighting, "It was reassuring that the system thoroughly confirmed its understanding aligned with my requests before proceeding to the next step."

However, there was also a comment—not directly associated with these two teams—that indicated "Bad point: there were too many confirmations." This suggests that confirmations should be minimized to maintain positive user experiences, possibly by only querying when the system's confidence level is low.

5.5.2 Speech Synthesis

In contrast to speech recognition, dialogues in which speech synthesis errors were pointed out[8] did not exhibit lower evaluation scores from either the Satisfaction or Completion perspective compared to dialogues without such errors (Table 5). This suggests an intriguing possibility: although speech synthesis errors might be memorable for

[7] Although evaluators could not see the speech recognition output in our data collection settings, they sometimes reported misrecognitions when they heard system responses that referred to unrelated words that sounded similar to those included in the evaluators' preceding utterances.

[8] The evaluators could not directly check the text being read aloud by the system, but they reported speech synthesis errors when they noticed clear mistakes in the synthesized speech, such as inappropriate readings of kanji characters in the system's responses.

interlocutors, they do not significantly affect the task performance or the user experience.

5.6 Analysis: Facial Expressions of Avatar

The preliminary round revealed some comments addressing issues related to the CG avatar's facial expressions, despite these expressions being directly unrelated to task completion. Considering the nature of the task, in which evaluators were required to choose tourist destinations, it was initially expected that their attention would predominantly focus on the Travel Viewer displaying detailed tourist information. Thus, it is particularly noteworthy that they may have allocated a certain degree of attention to the avatar's facial expressions.

Correlation analyses with Spearman's rank correlation coefficients conducted between the number of the avatar's facial expression changes in each dialogue and the evaluation scores revealed statistically significant positive correlations across all evaluation metrics from the Satisfaction perspective (Table 4). These results indicate that even in visually rich interactions, system behaviors not directly associated with task execution can affect interaction quality, highlighting the importance of efforts to develop systems with more natural and appealing behaviors.

5.7 Discussion

Based on the analyses presented thus far, two major findings have emerged.

First, most of the system behaviors that received a certain number of comments during the preliminary round were statistically confirmed to be key challenges affecting audio-visual task-oriented dialogue. This highlights the importance of conducting real interactions between humans and systems, gathering human feedback, and analyzing the feedback to uncover key challenges.

Second, merely transplanting methods or findings from prior research did not completely eliminate key challenges, particularly the response-delay and turn-taking problems analyzed in Section 5.2. This may be partly due to the increased task complexity and the richer information available to interlocutors in the current setting. As audio-visual task-oriented dialogue systems are expected to handle increasingly diverse and complex tasks in the future, it may become necessary to not only apply existing findings but also extend them in ways that account for the nature of each task.

6 Conclusion

In this paper, we presented an overview of the audio-visual task-oriented dialogue system competition we organized and the dataset constructed through this competition. To the best of our knowledge, this is the first publicly available large-scale dataset of audio-visual task-oriented dialogues between humans and systems. Based on feedback from evaluators, we identified key challenges in developing multimodal task-oriented dialogue systems that integrate both auditory and visual modalities. Our analysis further revealed that addressing some of the key challenges requires more than a direct application of existing research findings—it necessitates extending these insights, taking the nature of the task into account. We hope that the dataset and findings presented in this paper will contribute to more systematic and efficient research and development of audio-visual task-oriented dialogue systems.

Ethical Considerations

The dataset constructed in this study includes users' speech and facial images, necessitating careful consideration of privacy. We have obtained approval from the ethical review committee for departments at the Higashiyama Campus, Nagoya University, concerning data collection, usage, and publication.

Acknowledgments

This work was supported by JST Moonshot R&D Grant Number JPMJMS2011. We express our deepest gratitude to the participating teams in our competition for their valuable contributions. Our thanks also go to the interlocutors who took the stage in the final round to engage with the systems in front of the audience. We are also grateful to Professor Tetsuji Ogawa of Waseda University for arranging the venue for the final round. Finally, we are grateful to the audience members who attended the final round and provided third-party evaluations and survey answers.

References

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel Duckworth, Semih Yavuz, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. Taskmaster-1: Toward a Realistic and Diverse Dialog Dataset. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th*

International Joint Conference on Natural Language Processing, pages 4516–4525.

Olivier Canévet, Weipeng He, Petr Motlicek, and Jean-Marc Odobez. 2020. The MuMMER Data Set for Robot Perception in Multi-party HRI Scenarios. In *Proceedings of the 29th IEEE International Conference on Robot and Human Interactive Communication*, pages 1294–1300.

Jinsong Chen, Yuexin Zhang, and Luona Wang. 2025. The Impact of Service Robot Communication Style on Consumers' Continued Willingness to Use. *Collabra: Psychology*, 11(1):128020.

Yuya Chiba, Koh Mitsuda, Akinobu Lee, and Ryuichiro Higashinaka. 2024. The Remdis Toolkit: Building Advanced Real-time Multimodal Dialogue Systems with Incremental Processing and Large Language Models. In *Proceedings of the 14th International Workshop on Spoken Dialogue Systems Technology*, pages 1–6.

Lotte Eijk, Marlou Rasenberg, Flavia Arnese, Mark Blokpoel, Mark Dingemanse, Christian Doeller, Mirjam Ernestus, Judith Holler, Branka Milivojevic, Asli Özyürek, Wim Pouw, Iris Rooij, Herbert Schriefers, Ivan Toni, James Trujillo, and Sara Bögels. 2022. The CABB dataset: A multimodal corpus of communicative interactions for behavioural and neural analyses. *NeuroImage*, 264:119734.

Jan Gorisch, Corine Astésano, Ellen Gurman Bard, Brigitte Bigi, and Laurent Prévot. 2014. Aix Map Task corpus: The French multimodal corpus of task-oriented dialogue. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 2648–2652.

Akira Hayakawa, Saturnino Luz, Loredana Cerrato, and Nick Campbell. 2016. The ILMT-s2s corpus — A Multimodal Interlingual Map Task Corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 605–612.

Charles T. Hemphill, John J. Godfrey, and George R. Doddington. 1990. The ATIS spoken language systems pilot corpus. In *Proceedings of the Workshop on Speech and Natural Language*, page 96–101.

Matthew Henderson, Blaise Thomson, and Jason D Williams. 2014. The Third Dialog State Tracking Challenge. In *Proceedings of the 2014 IEEE Spoken Language Technology Workshop*, pages 324–329.

Ryuichiro Higashinaka, Michimasa Inaba, Zhiyang Qi, Yuta Sasaki, Kotaro Funakoshi, Shoji Moriya, Shiki Sato, Takashi Minato, Kurima Sakai, Tomo Funayama, Masato Komuro, Hiroyuki Nishikawa, Ryosaku Makino, Hirofumi Kikuchi, and Mayumi Usami. 2024. Dialogue System Live Competition Goes Multimodal: Analyzing the Effects of Multimodal Information in Situated Dialogue Systems. In *Proceedings of the 14th International Workshop on Spoken Dialogue Systems Technology*.

Yasuo Horiuchi, Yukiko Nakano, Hanae Koiso, Masato Ishizaki, Hiroyuki Suzuki, Michio Okada, Makiko Naka, Syun Tutiya, and Akira Ichikawa. 1999. The Design and Statistical Characterization of the Japanese Map Task Dialogue Corpus. *Journal of the Japanese Society for Artificial Intelligence*, 14(2):261–272.

Michimasa Inaba, Yuya Chiba, Zhiyang Qi, Ryuichiro Higashinaka, Kazunori Komatani, Yusuke Miyao, and Takayuki Nagai. 2024. Travel Agency Task Dialogue Corpus: A Multimodal Dataset with Age-Diverse Speakers. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(9):130.

Dinesh Babu Jayagopi, Samira Sheiki, David Klotz, Johannes Wienke, Jean-Marc Odobez, Sebastien Wrede, Vasil Khalidov, Laurent Nyugen, Britta Wrede, and Daniel Gatica-Perez. 2013. The vernissage corpus: a conversational human-robot-interaction dataset. In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction*, page 149–150.

Seokhwan Kim, Luis Fernando D'Haro, Rafael E. Banchs, Jason D. Williams, Matthew Henderson, and Koichiro Yoshino. 2016. The Fifth Dialog State Tracking Challenge. In *Proceedings of the 2016 IEEE Spoken Language Technology Workshop*, pages 511–517.

Seokhwan Kim, Yang Liu, Di Jin, Alexandros Papangelis, Karthik Gopalakrishnan, Behnam Hedayatnia, and Dilek Hakkani-Tur. 2021. "How robust r u?": Evaluating Task-Oriented Dialogue Systems on Spoken Conversations. *Preprint*, arXiv:2109.13489.

Seokhwan Kim, Yang Liu, Di Jin, Alexandros Papangelis, Behnam Hedayatnia, Karthik Gopalakrishnan, and Dilek Hakkani-Tür. 2022. Knowledge-grounded Task-oriented Dialogue Modeling on Spoken Conversations Track at DSTC10. In *Proceedings of the AAAI 2022 Workshop on Dialog System Technology Challenge*.

Kazunori Komatani and Shogo Okada. 2021. Multimodal Human-Agent Dialogue Corpus with Annotations at Utterance and Dialogue Levels. In *Proceedings of the 9th International Conference on Affective Computing and Intelligent Interaction*, pages 1–8.

Junyeong Kum and Myungho Lee. 2022. Can Gestural Filler Reduce User-Perceived Latency in Conversation with Digital Humans? *Applied Sciences*, 12(21):10972.

Takashi Minato, Ryuichiro Higashinaka, Kurima Sakai, Tomo Funayama, Hiromitsu Nishizaki, and Takayuki Nagai. 2023. Design of a competition specifically for spoken dialogue with a humanoid robot. *Advanced Robotics*, 37(21):1349–1363.

Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, and Erik Cambria. 2023. Recent advances in deep learning based dialogue systems: a systematic survey. *Artificial intelligence review*, 56(4):3055–3155.

Yuki Okamoto, Keisuke Imoto, Shinnosuke Takamichi, Ryosuke Yamanishi, Takahiro Fukumori, and Yoichi Yamashita. 2020. RWCP-SSD-Onomatopoeia: Onomatopoeic Word Dataset for Environmental Sound Synthesis. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop*, pages 125–129.

Zhenhui Peng, Kaixiang Mo, Xiaogang Zhu, Junlin Chen, Zhijun Chen, Qian Xu, and Xiaojuan Ma. 2020. Understanding User Perceptions of Robot's Delay, Voice Quality-Speed Trade-off and GUI during Conversation. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–8.

Ronald Poppe, Khiet P. Truong, and Dirk Heylen. 2011. Backchannels: Quantity, type and timing matters. In *Proceedings of the 11th International Conference on Intelligent Virtual Agents*, pages 228–239.

Libo Qin, Wenbo Pan, Qiguang Chen, Lizi Liao, Zhou Yu, Yue Zhang, Wanxiang Che, and Min Li. 2023. End-to-end Task-oriented Dialogue: A Survey of Tasks, Methods, and Future Directions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5925–5941.

Antoine Raux and Maxine Eskenazi. 2004. Non-Native Users in the Let's Go!! Spoken Dialogue System: Dealing with Linguistic Mismatch. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 217–224.

Amrita Saha, Mitesh Khapra, and Karthik Sankaranarayanan. 2018. Towards Building Large Scale Multimodal Domain-Aware Conversation Systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 696–704.

Shiki Sato, Yuta Sasaki, Shinji Iwata, Takato Yamazaki, Masato Komuro, Shoji Moriya, Masaya Ohagi, Hirofumi Kikuchi, Jie Yang, Asahi Hentona, Zhiyang Qi, Takashi Kodama, Akinobu Lee, Hiroyuki Nishikawa, Ryosaku Makino, Takashi Minato, Kurima Sakai, Tomo Funayama, Kotaro Funakoshi, Mayumi Usami, Michimasa Inaba, Tetsuro Takahashi, and Ryuichiro Higashinaka. 2025. The dialogue system live competition 7. *JSAI Technical Report, SIG-SLUD*, 103:01–08. (in Japanese).

Shuzheng Si, Wentao Ma, Haoyu Gao, Yuchuan Wu, Ting-En Lin, Yinpei Dai, Hangyu Li, Rui Yan, Fei Huang, and Yongbin Li. 2023. SpokenWOZ: A Large-Scale Speech-Text Benchmark for Spoken Task-Oriented Dialogue Agents. In *Proceedings of the 37th Conference on Neural Information Processing Systems*, pages 39088–39118.

Hagen Soltau, Izhak Shafran, Mingqiu Wang, Abhinav Rastogi, Wei Han, and Yuan Cao. 2023. DSTC-11: Speech Aware Task-Oriented Dialog Modeling Track. In *Proceedings of the Eleventh Dialog System Technology Challenge*, pages 226–234.

Philipp Spanger, Masaaki Yasuhara, Ryu Iida, Takenobu Tokunaga, Asuka Terai, and Naoko Kuriyama. 2012. REX-J: Japanese referring expression corpus of situated dialogs. *Language Resources and Evaluation*, 46(3):461–491.

Tetsuro Takahashi, Hirofumi Kikuchi, Jie Yang, Hiroyuki Nishikawa, Masato Komuro, Ryosaku Makino, Shiki Sato, Yuta Sasaki, Shinji Iwata, Asahi Hentona, Takato Yamazaki, Shoji Moriya, Masaya Ohagi, Zhiyang Qi, Takashi Kodama, Akinobu Lee, Takashi Minato, Kurima Sakai, Tomo Funayama, Kotaro Funakoshi, Mayumi Usami, Michimasa Inaba, and Ryuichiro Higashinaka. 2025. Analyzing dialogue system behavior in a specific situation requiring interpersonal consideration. In *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue*.

Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. 2020. Vision-and-dialog navigation. In *Proceedings of the Conference on Robot Learning*, pages 394–406.

Henry S. Thompson, Anne Anderson, Ellen Gurman Bard, Gwyneth Doherty-Sneddon, Alison Newlands, and Cathy Sotillo. 1993. The HCRC Map Task Corpus: Natural Dialogue for Speech Recognition. In *Proceedings of the Workshop on Human Language Technology*, page 25–30.

Mina Valizadeh and Natalie Parde. 2022. The AI Doctor Is In: A Survey of Task-Oriented Dialogue Systems for Healthcare Applications. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 6638–6660.

Marilyn A. Walker, Rebecca Passonneau, and Julie E. Boland. 2001. Quantitative and Qualitative Evaluation of Darpa Communicator Spoken Dialogue Systems. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 515–522.

Stephen Whittaker, Marilyn Walker, and Preetam Maloor. 2003. Should I Tell All?: An Experiment On Conciseness in Spoken Dialogue. In *Proceedings of the 8th European Conference on Speech Communication and Technology*, pages 1685–1688.

Jason Williams, Antoine Raux, Deepak Ramachandran, and Alan Black. 2013. The Dialog State Tracking Challenge. In *Proceedings of the 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 404–413.

Xuanming Zhang, Rahul Divekar, Rutuja Ubale, and Zhou Yu. 2023. GrounDialog: A Dataset for Repair and Grounding in Task-oriented Spoken Dialogues for Language Learning. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 300–314.

Team ID	N	Inf	Nat	App	Ple	Sat	Tru	Use	Rel	Ret	Exp	Avg	Suc	(Obj1+Obj2) / 2	Cmp
1	31	5.61	5.71	5.48	5.74	5.45	5.42	5.61	5.48	5.10	5.42	5.50	0.97	5.83	5.65
2	34	4.62	3.85	4.21	4.38	3.88	4.29	4.94	4.79	3.68	4.29	4.29	0.97	5.33	5.18
3	40	5.33	4.53	4.60	4.78	4.43	4.70	5.23	5.40	4.33	5.05	4.84	0.83	4.82	3.98
4	34	5.74	5.24	5.62	5.59	5.59	5.59	5.76	5.88	5.21	5.53	5.57	0.97	5.52	5.35
5	33	5.67	5.42	5.42	5.45	5.52	5.58	5.67	5.73	5.00	5.39	5.48	0.97	5.69	5.52
6	29	4.38	3.34	3.45	3.79	3.52	3.86	4.69	4.86	3.45	4.28	3.96	0.90	4.62	4.14
7	27	4.56	3.93	4.11	4.11	4.00	4.37	4.85	4.78	3.85	4.48	4.30	1.0	4.67	4.67
8	26	4.12	3.81	3.92	4.00	3.77	4.19	4.31	4.42	3.54	4.19	4.03	0.77	4.80	3.69
9	3	5.33	4.67	5.00	5.00	5.00	4.67	5.33	4.67	3.67	4.67	4.80	1.0	3.67	3.67

Table 6: Number of dialogues and averaged evaluation scores for each participating team in the preliminary round. N indicates the number of conversations obtained for each team, Avg indicates the satisfaction score (average of ten scores for satisfaction), and Cmp indicates the Completion score (Section 3.1). Note that the evaluation scores for dialogues that were excluded during the process of formatting the dataset have been excluded from the calculations in this table. Thus, the results for each team in the competition may differ slightly. In particular, the N value for Team ID 9 is notably small because its system frequently failed when interactions deviated from its predefined dialogue flow, and these failed dialogues were removed during the dataset curation process.

A System requirements for our competition

- The system must display the CG avatar using the specified software.[9]
- The system must be capable of communicating in Japanese.
- The system must utilize the specified text-to-speech API[10] for speech synthesis.
- Tourist spots must be selected from those registered in the provided "Rurubu DATA[11]" API.
- The system must employ the specified Travel Viewer, and all images displayed in the Travel Viewer must be obtained through the "Rurubu DATA" API.
- The system must initiate each dialogue.
- The determined tourist spot must be mentioned in the system's final utterance.
- Each dialogue must conclude within ten minutes.

B Demonstration of dialogues in the final round

To prevent systems from being evaluated solely based on exceptionally successful or unsuccessful dialogues, each system engaged in two dialogues (a total of six dialogues for the three systems), with the audience conducting third-party evaluations after each dialogue. The dialogues were conducted in two rounds. The first round consisted of one dialogue per system, conducted in a randomized order, followed by a second round using the same system order. The representative dialogue interlocutors consisted of six individuals: three counter sales staff from Japanese travel agencies who participated in the first round and three Japanese dialogue researchers who served as interlocutors in the second round.

C Details of Constructed Dataset

Table 6 shows the number of dialogues and evaluation scores for each team in the preliminary round.

[9] https://mmdagent-ex.dev.

[10] We specified Microsoft Azure's ja-JP-NanamiNeural. URL: https://learn.microsoft.com/azure/ai-services/speech-service/text-to-speech.

[11] API service provided by JTB Publishing, Inc. for acquiring and searching tourist attraction information. URL: https://solution.jtbpublishing.co.jp/service/domestic.