

Effects of Response Length on User Search Experience in Spoken Conversational Search

Ken Tobioka¹[0009–0001–7558–8297], Takehiro Yamamoto¹[0009–0001–7558–8297],
and Hiroaki Ohshima¹[0009–0001–7558–8297]

University of Hyogo, Kobe, Japan
af24h009@guh.u-hyogo.ac.jp
t.yamamoto@sis.u-hyogo.ac.jp
ohshima@ai.u-hyogo.ac.jp

Abstract. Response design plays a critical role in user experience in spoken conversational search (SCS). This study examines how two response styles—long response presentation (LRP) and short response presentation (SRP)—affect users’ cognitive load and search behavior under different task complexities. A user study ($N = 139$) was conducted using an LLM-based voice agent across low and high complexity search tasks. Results show that SRP reduced cognitive load and improved satisfaction, especially for low complexity tasks. However, in high complexity tasks, SRP led to more query issuance. This may indicate that users were actively engaging with the fragmented information to better understand the task or fill in missing details. These findings suggest that no single response style is optimal across all situations.

Keywords: SCS · exploratory search · user interface

1 Introduction

Voice-based conversational agents are gaining traction as natural language processing advances. With the rise of large language models (LLMs), such agents are expected to handle more complex tasks, including exploratory search tasks.

Long voice responses from LLMs often exceed users’ memory capacity, especially under high task complexity, making information processing difficult.

In this study, we compare two response presentation styles in spoken conversational search (SCS): **long response presentation (LRP)**, which presents information more comprehensively, and **short response presentation (SRP)**, which provides concise and focused responses. Our goal is to investigate how these styles affect users’ search experience.

Although prior work has examined response length preferences or compared visual and voice modalities [5,6,2], few studies have systematically investigated how response length in voice-only settings affects cognitive load, users’ subjective experience and search behavior.

To address this, we developed a desktop-based SCS system. We conducted a user study with 139 participants to examine the following research questions:

RQ1: Is the short response presentation method effective in spoken conversational search ?

RQ2: Do task complexity and response length affect users’ search behavior ?

Findings show that SRP reduced cognitive load and increased satisfaction in low complexity tasks, but triggered more queries in high complexity tasks. These results suggest that response strategies should adapt to task demands.

2 Related Work

Prior studies show that reading textual search results aloud can negatively affect the user experience [6]. Trippas et al. [5] found that users often prefer shorter summaries in voice-based interfaces, though task-dependent trade-offs exist between conciseness and informativeness.

For routine tasks, shorter responses improve usability [1], but this may not extend to exploratory search. Also, most prior work uses multimodal interfaces and does not isolate the effects of response length in voice-only settings [2] .

Our study extends this line of research by systematically examining how response presentations affect user experience and behavior under varying task complexities.

3 Methodology

3.1 Task Topics and Cognitive Complexity Levels

Two search scenarios were designed to reflect different levels of cognitive complexity, based on Trippas et al. [4].

- **Low complexity (Remember):** Where does cinnamon come from?
- **High complexity (Analyze):** Comparison of per capita alcohol consumption

3.2 Experimental Conditions

We defined two response styles:

- **LRP:** Comprehensive answers with multiple details, often in long sentences or bulleted form.
- **SRP:** Concise, one-to-two sentence summaries directly addressing the query.

Both styles used the gpt-4o-2024-08-06 model. Prompts were adjusted to encourage longer or shorter output.

Our system provided fully spoken interaction using OpenAI’s Whisper for speech-to-text conversion and a text-to-speech engine for audio output. Users initiated voice input by pressing an on-screen microphone button. Notably, system responses were presented exclusively via speech, without any textual display.

3.3 Flow of the User Study

After informed consent and a tutorial video ¹, participants completed a training task. Then they performed two main search tasks, each followed by questionnaires. The user study was approved by the research ethics committee at our university (No. UHIS-EC-2024-016).

3.4 Questionnaire

The questionnaire consisted of three parts: (1) a free-description item asking participants to summarize what they had understood through the search task, (2) the NASA Task Load Index (NASA-TLX) to measure cognitive load, and (3) participants rated the perceived effectiveness of response length in supporting information retrieval using a 5-point Likert scale (1: Strongly Disagree – 5: Strongly Agree).

3.5 Participants

We recruited 156 Japanese participants via Lancers.jp²; 139 completed the study successfully. Participants (aged 20–70s): 61.2% male, 38.8% female; 65.5% had prior experience with voice agents. Participants were compensated with approximately 1,100 JPY.

4 Results

We employed GLM to examine the main effects of response length (LRP vs. SRP) and topic complexity (low vs. high), as well as the interaction between these factors, on each NASA-TLX subscale, questionnaire (Q1–Q6) and number of issued queries and query issuance time.

4.1 Impact on Effectiveness (RQ1)

To investigate the effectiveness of SRP in SCS, we analyzed NASA-TLX scores as an indicator of cognitive load and participants' responses to Q1–Q6 as indicators of satisfaction with their search behavior. Table 1 shows the results.

Cognitive load: Mental demand showed significant main effects for response length ($F(1, 204) = 6.00, p < 0.05$) and task complexity ($F(1, 204) = 4.49, p < 0.05$). Overall workload showed significant main effects for both response length and task complexity, with a significant interaction ($F(1, 204) = 4.08, p < 0.05$). Post-hoc tests revealed SRP reduced workload especially under low complexity (mean: 54.34 for LRP vs. 44.55 for SRP, $t(121.9) = 3.10, p < 0.01$). This suggests that the benefit of **SRP** is more evident in **low complexity**

¹ <https://youtu.be/DSRzbFkI9f8>

² <https://www.lancers.jp/>

Table 1. Participant workload (NASA-TLX) and subjective ratings (Q1–Q6) across response length (LRP vs. SRP) and task complexity (low vs. high). Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Bold indicates better scores.

Measure		LRP		SRP		Response Length	Task complexity	Interaction
		Low	High	Low	High			
Cognitive load	Mental	44.17	45.82	32.62	43.21	*	*	0.07
	Temporal	22.93	25.54	15.63	23.87	0.09	*	0.47
	Frustration	24.51	28.25	17.62	27.49	0.17	*	0.31
	Overall	54.34	57.46	44.55	55.52	**	***	*
User satisfaction	Q1 (response length)	3.45	3.36	4.04	3.87	***	0.27	0.85
	Q2 (info helpfulness)	3.76	3.61	4.06	3.99	**	0.31	0.58
	Q3 (Info diversity)	4.15	3.87	3.87	4.01	0.51	0.51	*
	Q4 (Info coverage)	4.07	3.51	4.06	3.58	0.79	***	0.56
	Q5 (Opinion formation)	4.08	3.57	4.06	3.56	0.86	***	0.83
	Q6 (Info retention)	3.07	2.99	2.44	2.79	**	0.32	0.13

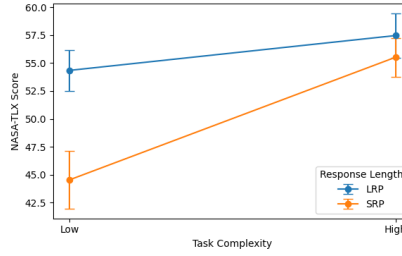


Fig. 1. Interaction effect between response length (LRP vs. SRP) and task complexity (low vs. high) on overall workload (error bars indicate the standard error of the mean).

tasks, while its advantage becomes less pronounced in high complexity tasks (see Figure 1).

User satisfaction: For Q1 and Q2, significant main effects of response length were observed, with **SRP** rated more appropriate ($F(1, 204) = 22.66$, $p < 0.001$) and helpful ($F(1, 204) = 10.06$, $p < 0.01$). Q3 showed an interaction effect ($F(1, 204) = 5.32$, $p < 0.05$), but follow-up t-tests showed only a weak trend favoring **LRP** in low complexity tasks. For Q6, **SRP** reduced the feeling of forgetting important information ($F(1, 204) = 9.67$, $p < 0.01$).

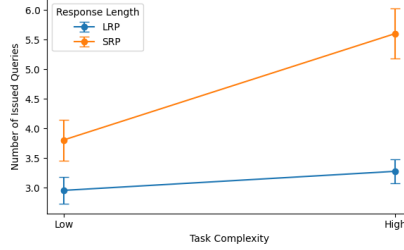
4.2 Impact on Search Behavior (RQ2)

To explore how the combination of response length and task complexity affects search behavior, we analyzed system interaction metrics including the number of issued queries and query issuance time. Table 2 shows the results.

Number of issued queries: We observed a significant interaction between response length and task complexity ($F(1, 204) = 5.54$, $p < 0.05$). Post-hoc t-tests confirmed the difference (mean: 5.55 for high vs. 3.85 for low, $t(134.33) = -3.12$, $p < 0.01$). This suggests that **SRP** prompts more iterative search in **high complexity tasks** (see Figure 2).

Table 2. Participant subjective ratings across response length (LRP vs. SRP) and task complexity (low vs. high). Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Measure		LRP		SRP		Response length	Task complexity	Interaction
		Low	High	Low	High			
Search behavior	Number of issued queries	2.96	3.28	3.80	5.60	***	***	*
	Query issuance Time (s)	11.89	9.19	13.90	13.44	*	0.38	0.30

**Fig. 2.** Interaction between response length (LRP vs. SRP) and task complexity (low vs. high) on the number of issued queries (error bars indicate the standard error of the mean).

Query issuance time: We observed a significant main effect of response length ($F(1, 141) = 4.21$, $p < 0.05$). This suggests that **LRP** may have provided more informative content, which helped users quickly identify follow-up questions and issue queries more efficiently.

5 Discussion

RQ1: SRP reduced cognitive load, particularly mental demand. Users found SRP responses appropriately concise (Q1), helpful for information acquisition (Q2), and easier to remember (Q6). A participant noted, “*The agent provided concise and direct answers, which was very helpful. When the responses were long, I tended to forget parts of the information.*”

However, a significant interaction in overall workload showed **SRP** was more effective only in **low complexity tasks**. In high complexity tasks, SRP’s brevity may have fragmented information, potentially increasing cognitive burden.

Q3 revealed a preference for **LRP** in low complexity tasks for acquiring diverse information, though some users found lengthy outputs excessive.

These findings suggest the need for *adaptive strategies*—e.g., providing LRP followed by a summary. One user suggested, “*It would be nice if the output adjusted to whether I want detailed or concise responses.*”

RQ2: Response length and task complexity affected behavior. **SRP** led to more queries, especially in **high complexity tasks**, suggesting users compensated for limited information. In contrast, **LRP** seemed to help users issue queries more efficiently—possibly due to richer content or added planning time

during longer responses. Systems might benefit from a dynamic approach—starting with LRP to aid exploration, then shifting to SRP to reduce overload.

Limitations: We did not assess the *quality* of acquired information (e.g., correctness or relevance). Future work could evaluate user summaries using information nuggets [3] and compare coverage with system outputs.

6 Conclusion

This study examined how response length and task complexity affect user experience in SCS. The main findings are twofold. First, SRP significantly reduced cognitive load and increased user satisfaction, particularly in low complexity tasks. SRP is likely effective in mitigating the memory burden inherent in SCS. Second, when the task complexity increased, SRP condition led to a significantly higher number of issued queries, reflecting more active search behavior. However, this may also increase cognitive strain, as users need to reconstruct fragmented information across multiple interactions. This suggests the need for task-adaptive rather than uniform response strategies.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers JP24K03228, JP22H03905, JP21H03554, and JP21H03775.

References

1. Haas, G., Rietzler, M., Jones, M., Rukzio, E.: Keep it short: A comparison of voice assistants' response behavior. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. pp. 1–12 (2022)
2. Kaushik, A., Jones, G.: Examining the potential for conversational exploratory search using a smart speaker digital assistant. In: Proceedings of the 7th International Conference on Human Computer Interaction Theory and Applications. pp. 305–317 (2023)
3. Lin, J., Demner-Fushman, D.: Will pyramids built of nuggets topple over? In: Proceedings of the Human Language Technology Conference of the NAACL, Main Conference. pp. 383–390. Association for Computational Linguistics (2006)
4. Trippas, J.R., Spina, D., Cavedon, L., Sanderson, M.: How do people interact in conversational speech-only search tasks: A preliminary analysis. In: Proceedings of the 2017 Conference on Human Information Interaction and Retrieval. pp. 325–328 (2017)
5. Trippas, J.R., Spina, D., Sanderson, M., Cavedon, L.: Towards understanding the impact of length in web search result summaries over a speech-only communication channel. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 991–994 (2015)
6. Vtyurina, A., Clarke, C.L.A., Law, E., Trippas, J.R., Bota, H.: A mixed-method analysis of text and audio search interfaces with varying task complexity. In: Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval. pp. 61–68 (2020)