Gesture and Audio-Haptic Guidance Techniques to Direct Conversations with Intelligent Voice Interfaces

Shwetha Rajaram*
Reality Labs Research, Meta
Toronto, Ontario, Canada
shwethar@umich.edu

Hemant Bhaskar Surale
Reality Labs Research, Meta
Toronto, Ontario, Canada
hemantsurale@meta.com

Codie McConkey
Reality Labs Research, Meta
Toronto, Ontario, Canada
codiemcconkey@meta.com

Carine Rognon
Reality Labs Research, Meta
Redmond, Washington, USA
carinerognon@meta.com

Hrim Mehta
Reality Labs Research, Meta
Toronto, Ontario, Canada
hrimrmehta@meta.com

Michael Glueck
Reality Labs Research, Meta
Toronto, Ontario, Canada
mglueck@meta.com

Christopher Collins
Reality Labs Research, Meta
Toronto, Ontario, Canada
chriscollins@meta.com

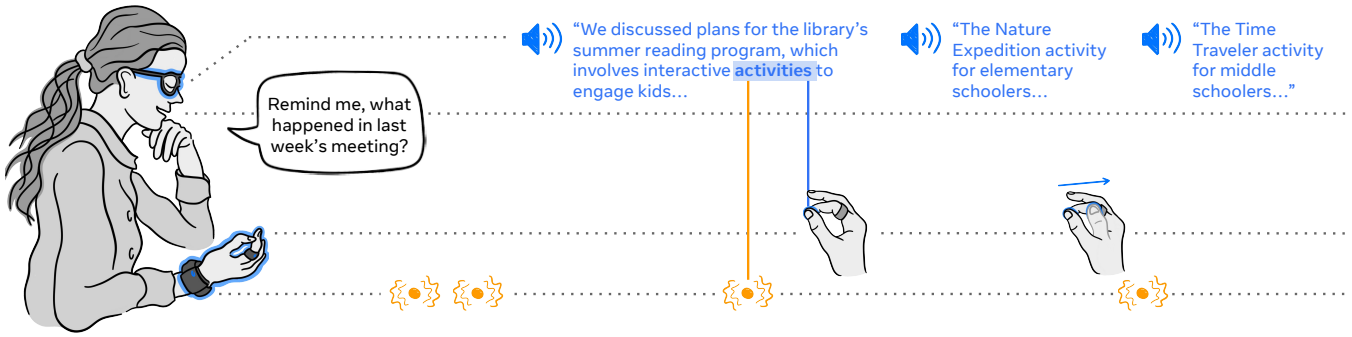

Figure 1: To reduce friction in conversations with intelligent voice interfaces, this work develops gesture and audio-haptic interaction techniques that allow users to rapidly navigate and manage the timing of conversations.

Abstract

Advances in large language models (LLMs) empower new interactive capabilities for wearable voice interfaces, yet traditional voice-and-audio I/O techniques limit users' ability to flexibly navigate information and manage timing for complex conversational tasks. We developed a suite of gesture and audio-haptic guidance techniques that enable users to control conversation flows and maintain awareness of possible future actions, while simultaneously contributing and receiving conversation content through voice and audio. A 14-participant exploratory study compared our parallelized I/O techniques to a baseline of voice-only interaction. The results demonstrate the efficiency of gestures and haptics for information access, while allowing system speech to be redirected and interrupted in a socially acceptable manner. The techniques also raised user awareness of how to leverage intelligent capabilities. Our findings inform design recommendations to facilitate role-based collaboration between multimodal I/O techniques and reduce users' perception of time pressure when interleaving interactions with system speech.

CCS Concepts

• **Human-centered computing** → **Natural language interfaces**; *Gestural input*; *Haptic devices.*

Keywords

multimodal interaction, voice interfaces

*This work was done while the author was a Research Intern at Meta. The author is also affiliated with the University of Michigan School of Information.

CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3714310

ACM Reference Format:
Shwetha Rajaram, Hemant Bhaskar Surale, Codie McConkey, Carine Rognon, Hrim Mehta, Michael Glueck, and Christopher Collins. 2025. Gesture and Audio-Haptic Guidance Techniques to Direct Conversations with Intelligent Voice Interfaces. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3706598.3714310

1 Introduction

Advances in large language models (LLMs) are enabling new opportunities for wearable voice interfaces that let users perform complex tasks without visual displays. While traditional deterministic voice assistants are typically used for short, low-risk transactions (e.g., factual questions, reminders, weather updates [32, 43]), the HCI community is exploring how new LLM-powered display-free devices, such as the Ray-Ban Meta Smart Glasses[1] or Amazon Echo Frames[2], can facilitate more in-depth conversations to provide context-aware activity guidance [31], document daily events [7], or write emails on-the-go [64].

However, these new capabilities can exacerbate existing challenges with managing conversations with voice interfaces, due to the expertise and overhead required to interact with today's LLMs. The linear nature of voice input and audio output limits users' ability to control the *order* of system responses, e.g., to branch into threads of discussion for sensemaking tasks [51, 52] or backtrack to correct system errors [26, 33]. Managing the *timing* of conversations also poses difficulties, as lengthy or irrelevant system responses can lead to unpredictable time investments [15, 32, 55]; users must carefully time their speaking and silence to distinguish between continuing and new requests [43]. Addressing these challenges with *navigational* and *temporal flow* is critical to enabling users to fully leverage advanced LLM-enabled functionality.

To empower users to flexibly direct conversations with intelligent voice interfaces, **we developed a suite of gesture and audio-haptic guidance techniques to parallelize interactions related to conversation *content* and *control***. This approach draws inspiration from prior systems that structure a role-based collaboration between multimodal input techniques (e.g., Pen+Touch [16], BISHARE [65]). We reserve voice and audio for specifying details in a conversation, while introducing gestures and haptics for managing flow (e.g., adjusting response conciseness via gestures) and receiving interaction guidance (e.g., audio-haptic nudges highlighting topics to dive deeper into). To inform the design of these parallelized I/O techniques, we reviewed and classified prior examples of multimodal interaction with conversational UIs, including both visual [25, 33] and audio-only interfaces [21, 59].

In a 14-participant exploratory study, we compared our gesture and audio-haptic guidance techniques to a voice-only baseline, investigating their effectiveness for managing navigational and temporal flow. To enable this, we developed an LLM-enabled voice assistant integrated with Ray-Ban Meta Glasses for audio and wearable devices for gestures and haptics (Fig. 1, 10). Participants found that our gesture and haptic-driven techniques enabled more efficient access to information, offering socially acceptable mechanisms to interrupt and redirect system speech to align with their goals. The techniques also promoted their awareness of possible future actions in the conversation. However, haptic cues interleaved with system speech increased participants' perception of time pressure, and they tended to overlook natural voice-driven interactions when gestures were available. We discuss design improvements to mitigate these tradeoffs in future work.

[1]**Ray-Ban Meta Smart Glasses:** https://www.meta.com/smart-glasses
[2]**Amazon Echo Frames:** https://www.aboutamazon.com/news/devices/introducing-next-generation-echo-frames-carrera-smart-glasses-with-alexa

Our work contributes: *(1)* a suite of multimodal interaction techniques (combining voice, gestures, audio, and haptics) to facilitate conversations with voice interfaces, instantiated in a functional system; *(2)* an exploratory study demonstrating the benefits of our techniques for rapid access to information and flexible navigation within conversations; *(3)* design recommendations to better facilitate role-based collaboration between the I/O modalities.

2 Related Work

Our research extends prior work on conversational interfaces and multimodal interaction.

2.1 Challenges with Conversational Interaction

Conversational user interfaces are widely adopted for their natural interaction style and expressive power; however, they exhibit key interaction challenges that lead even experienced users to restrict their usage to short and low-risk tasks (e.g., daily reminders, playful or humorous interactions) [32, 43]. Our work seeks to improve the **navigational flow** (the *order* of user input and system output) and **temporal flow** (the *timing* of user input and system output) for the next-generation of display-free, AI-enabled voice interfaces, such as the Ray-Ban Meta Glasses or Amazon Echo Frames. As such, we review prior empirical studies of both audio-only conversational interfaces (e.g., smart speakers [32, 43]) and LLM-enabled visual interfaces (e.g., ChatGPT [26]).

Challenges with navigational flow. Because voice input and audio output are delivered through low bandwidth channels, users are constrained to conversing in a linear manner rather than flexibly navigating system responses, including *(1)* branching into threads of discussion (e.g., for sensemaking tasks [51, 52]) and *(2)* backtracking to clarify intent (e.g., due to errors in speech recognition or misaligned LLM responses [26, 33]). Additionally, using audio as the sole output channel results in interaction guidance (e.g., suggestions for follow-up questions) being interleaved with the current conversation content. This forces users to context switch between processing the details of the conversation and determining appropriate next steps to continue the interaction.

Challenges with temporal flow. Today's voice UIs have limited activation periods to avoid always-on speech transcription, requiring users to carefully their speech and intentional silence to either maintain the context window in the existing request or initiate a new request [1, 43, 58]. Especially for complex tasks (e.g., formulating recommendations or evidence-based explanations [5]), LLMs tend to generate lengthy responses that are difficult to parse and remember through an audio-only modality [15, 18].

Beyond managing timing within a conversation, the overall time investment is often unpredictable. With traditional deterministic voice UIs, non-expert users employ repetitive question-answering strategies to gain an understanding of the floor and ceiling of voice UIs' capabilities [32]. This back-and-forth alignment process may be further prolonged with LLM-enabled functionality, which requires more effort from users to adapt their expectations to the less transparent decision-making processes of LLMs [55].

Addressing these challenges is critical to enabling end-users to leverage voice interfaces for complex conversational tasks beyond the simple, transactional tasks they are currently used for [32, 43].

To this end, our research develops multimodal interaction techniques (combining voice, gestures, audio, and haptics) for users to flexibly direct conversation flows. We discuss considerations for multimodal design in the next section.

2.2 Multimodal Information Processing

HCI research has long studied how leveraging a broad range of humans' sensorimotor capabilities can enhance users' performance when using digital interfaces [12]. From a cognitive psychology perspective, multimodal interaction is feasible due to humans' ability to process multiple sensory inputs in parallel, using distinct working memory subsystems–such as the visual-spatial and verbal-phonological loops–to overcome limitations of individual processing loops [3, 8, 30]. Specific to conversational UIs, prior studies have found multimodal interaction to support increased efficiency [14, 36, 39], expressive communication and information transfer in collaborative scenarios [13, 27], information retention [9], and richer engagement with system content [49].

However, to enable these benefits, interactions need to be carefully designed to integrate multimodal input (*fusion*) and deliver multimodal output (*fission*) while minimizing cognitive load [44, 50]. A popular input design pattern is *decision-level fusion* [50]: combining multiple input techniques at a semantic level to enable role-based collaboration [16, 37, 53, 65]. For example, Pen+Touch leveraged a digital pen for writing and multi-touch gestures for manipulating content in digital workspaces [16]. XDBrowser [37] explored role-based cross-device patterns for large and small mobile devices, e.g., to extend displays or separate viewing and editing interfaces. To reduce friction in interactions with intelligent voice interfaces, our work takes a similar approach to parallelize actions related to conversation content (via voice and audio) and conversation control (via gestures and audio-haptic cues).

To deliver multimodal output, prior work explored *synchronizing* content across modalities (e.g., notifying users through both audio and haptics) to allow users to cross-reference information and reinforce their understanding, particularly in learning scenarios [8, 12]. Conversely *stratifying* output across modalities enables multi-tasking (e.g., delivering silent visual notifications while the user listens to a podcast), but should be designed to avoid splitting users' attention across multiple salient output channels, which can hinder information retention [8].

Our work builds on these principles to develop a suite of gesture and audio-haptic guidance techniques, working alongside voice and audio I/O to facilitate conversations with voice interfaces. Through a study with 14 end-users, we investigated the impact of both synchronized techniques (e.g., simultaneous audio-haptic cues to confirm users' gestures) and stratified techniques (e.g., haptics interleaved with the system's response to emphasize interaction opportunities) on participants' conversation flows.

2.3 Multimodal Interaction with Voice Interfaces

Our work extends HCI systems research applying multimodal design principles to enable more expressive interactions with voice interfaces. Seminal works such as Put-that-there [6] and QuickSet [10] combine pointing and voice commands for mixed-initiative manipulation of digital content. In recent systems, we observe similar gesture-based techniques for referencing and manipulating physical objects: GazePointAR [29] leverages pointing to disambiguate users' references when querying real-world information; Minuet [22] enables multimodal control over IoT devices; Wu et al.'s design space [59] demonstrates full-body input techniques to voice interfaces.

	Expressiveness	Time Cost
Voice input	high	high
Gesture input	med	low
Audio output	high	high
Haptic output	med	low

Figure 2: I/O Modality Properties. Our multimodal interaction techniques capitalize on voice input and audio output for their high expressiveness. To mitigate challenges with the high time cost of processing audio, we incorporate gestures and haptics as highly efficient interaction modalities.

While prior systems primarily leverage gestures for atomic or transactional tasks (e.g., ordering food [59], changing temperature [22]), we also observe trends towards enabling complex text-based tasks that traditionally require visual interfaces. For example, Earpod [63] and AudioHallway [46] support silent navigation of audio files through mouse gestures or head movement. GlassMail [64] and Voice+Tactile [21] enable coarse- and fine-grained text editing in on-the-go scenarios using voice and touch input, respectively. Similarly, VERSE [56] integrates voice and touch input to support information-seeking tasks with screen readers for blind and low-vision users.

Building on these works, we developed multimodal interaction techniques that not only enable users to guide system speech through gestures, but also provide haptic channels to raise users' awareness of system state and possible future actions. We focus on new types of conversational tasks that are made possible by LLMs and require more deliberation and turn-taking, e.g., seeking evidence-based explanations, debating conflicting perspectives, or crafting recommendations [5].

3 Parallelized I/O Techniques for Conversing with Voice Interfaces

Given that key challenges with voice interfaces stem from the linear nature of voice input and audio output, we explored how to parallelize interactions related to conversation *content* and *control* through developing a suite of gesture and audio-haptic guidance techniques. As shown in Figure 2, gestures and haptics are less expressive than voice and audio (i.e., they convey less rich or nuanced information [20]), but are more efficient for users to perform and process [6, 54]. These properties make gestures and haptics well-suited for rapid functions to control conversation flow (e.g., adjusting conciseness via gestures or receiving interaction guidance through audio-haptic cues), while voice and audio can be reserved for posing prompts and receiving responses (Fig. 3).

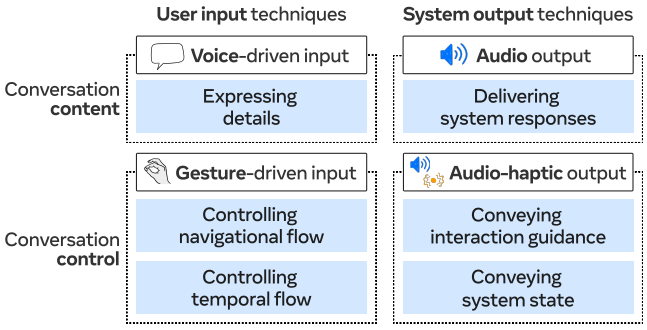

Figure 3: Role-Based Collaboration between I/O Modalities. Traditional voice interfaces leverage voice input and audio output to exchange conversation *content*. Our multimodal interaction techniques parallelize interactions related to conversation *control* through gesture input (to manage navigational and temporal flow) and audio-haptic output (to convey interaction guidance or system state).

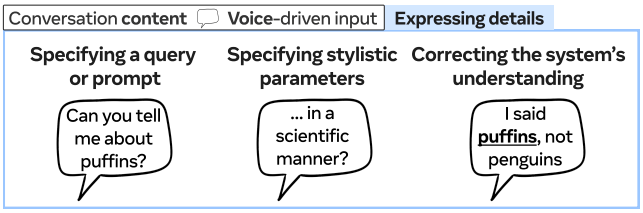

Figure 4: Voice to Express Details.

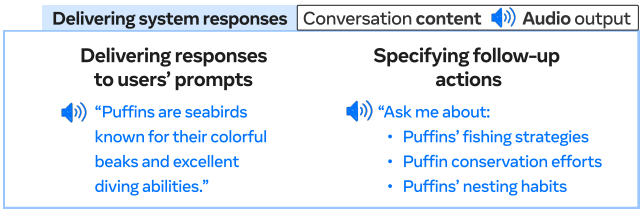

Figure 5: Audio to Deliver System Responses.

To understand the range of implicit and explicit interactions to parallelize, we classified gesture and haptic-based techniques for conversing with voice interfaces, based on our own prototypes and a literature review of multimodal interaction with conversational UIs, including both visual interfaces (e.g., chatbots) and audio-only interfaces. Our work builds on prior design spaces of gesture-driven, task-oriented interactions with voice UIs (e.g., controlling IoT devices, placing orders) [21, 22, 59], but instead we focus on facilitating complex human-AI conversations (e.g., seeking evidence-based explanations or debating conflicting perspectives [5]).

While the latest generation of wearable AI devices, such as the Ray-Ban Meta glasses, incorporate camera data (e.g., for world-querying tasks [29, 31]), we scoped our exploration to voice and gesture input to conduct a more focused comparison to traditional voice-driven interaction. We also note that other multimodal inference techniques are possible (e.g., *feature-level fusion* [50] combining gesture detection and gesture speed to infer users' preferences for the pace of conversation) and would be interesting to explore in future work.

Figures 4-8 depict our suite of parallelized I/O techniques informed by our literature review; techniques in bold are ones we implemented (Sec. 4) and investigated through an exploratory study with end-users (Sec. 5).

3.1 Traditional I/O for Conversation Content

A wide range of interactive systems (e.g., wearable assistants, smart home devices, mixed reality experiences) incorporate voice and audio to offer users natural interaction techniques with high expressive leverage. In commercial AI-enabled voice interfaces (e.g., ChatGPT's Advanced Voice Mode[3]) and research systems [31, 59, 64], voice input is used to express details in a conversation: specifying queries or prompts, specifying stylistic qualities, or correcting system errors (Fig. 4).

[3]ChatGPT's Advanced Voice Mode: https://help.openai.com/en/articles/9617425-advanced-voice-mode-faq

Similarly, the expressiveness of audio makes it an ideal modality to deliver system responses to the user or issue interaction guidance that requires significant explanation, such as specifying follow-up topics to further explore in a conversation (Fig. 5).

Trade-offs: Voice and audio are transmitted through low bandwidth channels, requiring users to speak and listen sequentially to convey and consume information. As discussed in Section 2.1, these tradeoffs pose challenges for managing navigation and timing within conversations, e.g., difficulties backtracking to correct errors [26, 33] and processing information in lengthy responses [15].

Next, we describe how we envision mitigating these challenges by using gestures, a more efficient input modality, to guide conversation flows alongside voice and audio I/O.

3.2 Gesture Input for Conversation Control

Prior conversational UIs place gestures in supporting roles to speed up intent specification (e.g., pointing to disambiguate vague language [29], performing hand poses to add numerical information to voice directives [59]). Among recent visual interfaces for conversing with LLMs, we also see trends towards gesture-driven direct manipulation to iteratively refine prompts [25, 33] or identify future directions for discussion [18, 52].

Inspired by these prior examples, we developed a set of **gesture-driven techniques to flexibly navigate system responses** (Fig. 6):

(1) **Selecting follow-up topics:** To redirect the conversation based on system-suggested, verbally-read options, users can tap to **`Select`** a topic to explore further.
(2) **Going deeper into topics:** Tapping while the system is speaking drives the conversation toward the topic just discussed. However, compared to visual interfaces [33, 52], directly manipulating audio responses is more challenging, as the linear nature of audio requires precisely timing interactions. To help users understand when and how they can manipulate system speech, we incorporate haptic nudges to emphasize keywords to **`Go Deeper`** into (Sec. 3.3).

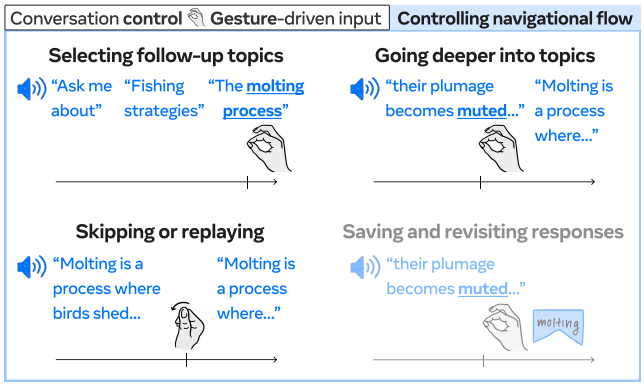

Figure 6: Gestures to Control Navigational Flow.

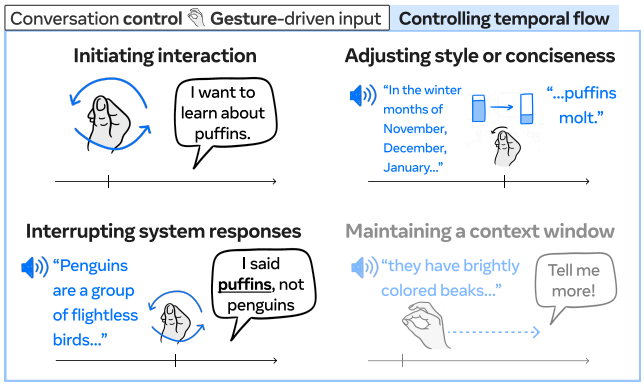

Figure 7: Gestures to Control Temporal Flow.

(3) **Skipping or replaying system responses:** Inspired by lightweight media controls in audio-only interfaces [41, 63]), we allow users to `Swipe Left` to **`Replay`** the current response and `Swipe Right` to **`Skip`** to the next sentence.
(4) **Saving system responses:** While not implemented, we envision bookmarking topics for later reference via gestures interleaved with system speech.

We also explored four classes of techniques to **directly manipulate the timing of conversations** (Fig. 7):

(1) **Initiating interaction with voice interfaces:** Aligned with prior systems [41, 59], we use a distinctive gesture (a quick wrist rotation) to **`Wake`** the voice interface in a subtler manner than using a verbal wake word.
(2) **Adjusting the style or conciseness of system responses:** We implemented three techniques to help users tailor the response length and level of detail to their needs. First, an interruptive technique: **`Restyle Response`** allows users to gesture and verbally provide stylistic instructions for the system to regenerate its current response. Next, two non-interruptive techniques: **`Wrap it Up`** ends the response gracefully at the next sentence, and **`Tell me More`** extends the current response with additional details. Aligned with our work, prior systems implemented techniques for keyword summarization to reduce users' audio processing effort [21, 64] and progressive topic exploration, using link navigation to traverse knowledge graphs [56].
(3) **Interrupting system responses** [59]: To redirect conversations in a more socially acceptable way than verbally interjecting, the `Wake` gesture can also be used interruptively, stopping system speech and activating the mic for the user to speak.
(4) **Maintaining the context window of the conversation:** In traditional voice UIs, silence signals the end of a conversation, leading the system to clear its conversation history. To enable users to pause and process information, we envision users gesturing to "hold their place" in the conversation. Minuet [22] proposed maintaining context via repeated gestures (e.g., to tell a smart thermostat to keep increasing the temperature).

Trade-offs and Design Considerations: A one-to-one mapping between gestures and system functions can pose **memorability** challenges [34]. To reduce the number of distinct gestures, our implementation differentiates gestures performed while the user is talking, while the system is talking, or during silence. This enables a one-to-many mapping (e.g., `Pinch` gestures correspond to `Pause` while the system is talking and `Keep Listening` while the user is talking). Future approaches could contextually adapt gesture mappings based on frequent actions for a particular task. For example, a `Swipe` gesture could translate to `Skip` for a summarization task or `Tell me More` when learning about new topics.

We also note that gesture-driven interaction may conflict with a core advantage of voice interfaces: enabling **hands-free task completion** [32]. While we prototyped our interaction techniques with one-handed gestures (Sec. 4), we envision the techniques translating to other forms of gesture recognition (e.g., microgestures designed to work under hand pose or location constraints [19]).

3.3 Audio-Haptic Output for Conversation Control

The previous section demonstrated user-driven interaction techniques combining voice and gestures to guide system responses. Next, we present audio-haptic interactions for the system to guide users to *(1)* understand system state and *(2)* plan future interactions.

Traditional voice interfaces use non-verbal audio cues to notify users about system status without a visual display (e.g., sound effects to indicate loading). However, with audio-only conversational UIs, there is a risk that the simultaneous processing of both verbal and non-verbal audio could overload users' auditory perception capabilities [12, 40]. To parallelize system output related to conversation content and interaction guidance, we leverage vibrotactile haptics to repeat ambient interaction cues through a separate sensory channel, which has been shown to reduce user error [48].

A wide range of digital devices employ haptics to provide ambient awareness and interaction guidance, including mobile phones, smartwatches, and mixed reality systems [2, 4]. Haptics are emerging in conversational AI systems as well, e.g., ChatGPT's mobile app incorporates typing haptics to notify users about response loading time and length.

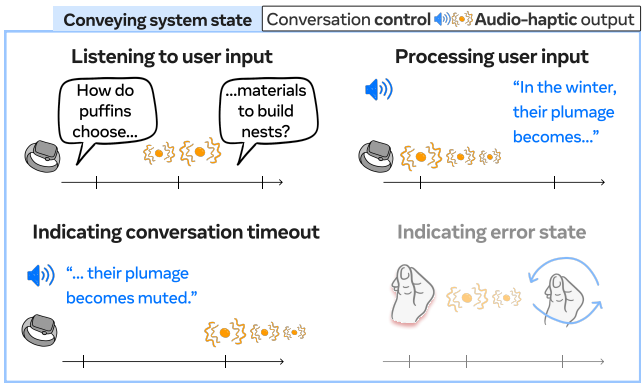

Figure 8: Audio-Haptic Cues to Convey System State.

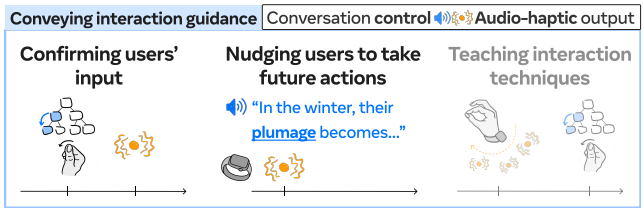

Figure 9: Audio-Haptic Cues to Convey Interaction Guidance.

First, to **reinforce users' awareness of the current status of the conversation**, we designed a set of audio-haptic indicators (Fig. 8) for the following system states:

(1) **Listening to user input:** Inspired by prior conversational UIs and long-distance communication systems [21, 24], our implementation issues vibrotactile feedback to assure users that their speech is being captured. We trigger this haptic feedback when users are speaking longer prompts or holding a **Keep Listening** gesture.
(2) **Processing user input:** Aligned with traditional voice UIs, we also deliver audio-haptic cues to indicate system loading times.
(3) **Indicating conversation timeout:** Silence typically signals the end of a conversation to voice interfaces. To help users anticipate when the system will forget their previous interaction, we issue subtle haptic feedback after 5s of silence. Speaking within another 5s allows users to continue the conversation, while remaining silent clears the history. Prior systems explored salient haptic feedback to achieve the opposite effect: encouraging users to reduce screen time [38].
(4) **Indicating error state:** While not implemented, we envision audio-haptic cues to notify users of various system faults (e.g., gesture recognition errors, LLM response failures [45]).

Finally, we explored a set of techniques to **support awareness of potential future interactions and guide users on how to execute them** (Fig. 9).

(1) **Confirming users' input:** Our implementation issues audio-haptic cues for various types of gestural input to assure users that the system is acting upon their intent. Based on our pilot studies, we decided to use distinct cues for only two system functions: microphone activation and deactivation. For all other gestures (e.g., `Go Deeper`, `Wrap it Up`), we deliver a uniform confirmation signal to reduce users' cognitive load when distinguishing between cues.
(2) **Nudging users to take future actions:** Inspired by spatial cues in audio browsing interfaces [35, 46], we designed ambient haptic patterns to raise users' awareness of *when* to execute interactions. Haptic nudges interleaved with system speech indicate keywords for users to explore more deeply (paired with the `Go Deeper` gesture).
(3) **Teaching users to perform interaction techniques:** In a future implementation, we also envision cues to teach users *how* to execute interactions. To this end, Voice+Tactile provides finger-level gesture guidance via a touchpad tactile interface [21] and Xu et al.'s haptic patterns [60] convey semantic associations (e.g., rendering feedback around a wristband to simulate grabbing for a `Save` function).

Trade-offs and Design Considerations: Transmitting meaningful information through haptics is a known design challenge, while distinguishing unique haptic patterns poses challenges for users [54]. Our pilot studies suggested that audio counterparts to haptics were essential for participants to understand system state (e.g., mic activation, system loading). An exception to this rule was cues synchronized with the voice interface's speech (e.g., keyword nudges), for which participants preferred haptic-only feedback to avoid distracting from the audio response.

4 Implementation

To demonstrate interaction techniques along our design space, we developed an LLM-enabled voice interface as a Unity application integrated with two in-house research devices for gesture input and haptic feedback (Fig. 10). We simulated a smart glasses form factor by streaming audio from the Unity client to Ray-Ban Meta glasses via Bluetooth.

LLM-enabled Voice Interface (Fig. 10A): To power the intelligent voice interface, we used GPT-4o to develop 4 agents that generate, parse, or modify various aspects of the conversation. The primary agent is an `Informational Audio Assistant` that generates responses to users' prompts given a truncated conversation history and optional stylistic instructions. We embedded prior work's design guidelines for voice interfaces [28, 57] into the GPT-4o prompts to improve the quality of responses (e.g., to encourage conciseness and predictable sentence structures). We also developed a `Keyword Identifier` that highlights important words in the system's response, a `Follow-up Generator` that suggests subtopics for users to explore, and a `Response Extender` that generates more details at the end of the system's current response. Appendix A.3 shows the prompts for all agents.

We used the Meta Voice SDK[4] (version 66.0) for *dictation* (to convert users' speech into a text-based request for GPT-4o), *text-to-speech* (to translate GPT-4o's text responses into audio output for the user), and *voice commands* (to activate the microphone via a wake word in our user study, as described in Sec. 5). To enable interleaving gesture input and haptic feedback with the system's

[4]**Meta Voice SDK:** https://developer.oculus.com/downloads/package/meta-voice-sdk/

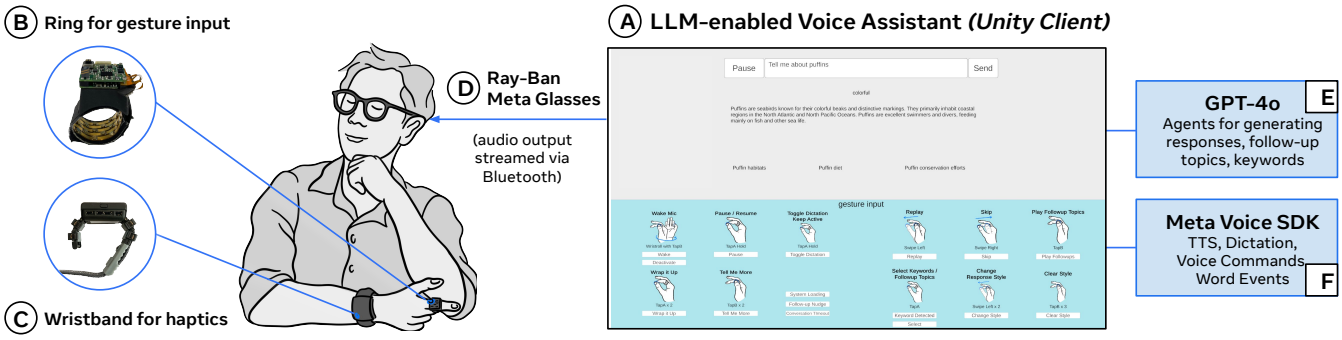

Figure 10: Voice Interface System Architecture. We developed the LLM-enabled voice assistant as a Unity application *(A)* using GPT-4o *(E)* and the Meta Voice SDK *(F)*. We connected the Ray-Ban Meta Glasses as a Bluetooth device to deliver audio output *(D)* and used two in-house, wireless wearables for gesture input *(B)* and haptic feedback *(C)*.

speech, we used *text-to-speech word events*, which provide timing offsets for every word in an audio clip.

Gesture input device (Fig. 10B): We used an in-house wearable, wireless ring that uses active electrical sensing to detect contact between the user's index finger and other parts of their skin. The ring is instrumented with a set of electrodes that measure impedance changes that occur when the index finger touches skin (similar to [23, 61, 62]), an IMU that infers the orientation of the finger, and an optical proximity sensor that differentiates across multiple touch locations. Using this setup, we distinguish between two tap positions: the distal phalanx and the proximal phalanx, which we refer to as *A* and *B* herein. A heuristics-based signal processing pipeline fuses data from these sensors to infer on-skin touch events and one-handed gestures, such as thumb-tap on position A or B. These gesture events that are streamed to the Unity client.

For our user study, we designed a gesture mapping that leverages simple semantic metaphors to help users remember which gestures correspond to system functions (Fig. 12). For example, Tap B sequences correspond to going forwards in the conversation (`Skip` and `Tell me More`), while Tap A sequences signal backtracking or stalling the conversation (`Replay` and `Wrap it Up`). While we instantiate the gesture set with our ring device, we designed and developed the interaction techniques to be agnostic to the gesture recognition method: the voice interface could be extended to accept gesture events via keyboard simulation or camera-based gesture input detected via cameras.

Haptic device and audio-haptic cues (Fig. 10C): To enable haptic feedback, we used another in-house wearable device: a wireless wristband lined with 4 vibrotactile actuators, which are evenly spaced on the top, bottom, left, and right of the wrist. We worked with a professional sound designer in our organization to create audio cues for three categories: *(1)* confirming gesture detection, *(2)* supporting awareness of system state (when the system is loading responses or listening to users' speech), and *(3)* nudging users to take future actions (emphasizing keywords or encouraging the user to continue the conversation after periods of silence). We then created haptic patterns matching the frequency and amplitude of the audio cues (which are played through all 4 actuators on the wristband simultaneously). We designed these audio-haptic cues to be just noticeable to prevent interrupting the flow of conversation.

5 Exploratory User Study

To investigate how introducing gestures and haptics as parallel I/O channels for voice-based interaction impacts the navigational and temporal flow of conversations, we conducted an exploratory study with 14 end-users of voice interfaces. Through two conversational tasks with our LLM-enabled voice interface (Sec. 4), participants compared traditional voice-only interaction with our gesture and haptic-driven guidance techniques.

The study was approved by the Institutional Review Board (IRB #00000971). We conducted 90-minute in-person study sessions. Participants were compensated for their time.

5.1 Participants

We advertised the study via public mailing lists, contacting a random distribution of 50 interested individuals. Participants met inclusion criteria requiring English fluency, hearing capabilities, and the ability to perform hand gestures. Since we used Ray-Ban Meta Glasses as the voice interface form factor, we recruited participants who would be comfortable removing their glasses to ensure proper fit; vision impairment was not a disqualifier. From the 16 completed sessions, we omitted 2 participants from our analysis, one due to a temporary *text-to-speech* failure with the Meta Voice SDK and the other due to low performance on our attention check questions.

The final sample of 14 participants had an average age of 31 years (range: 21-61), with 5 women, 7 men, and 2 participants preferring not to report.

Experience with voice-based interfaces: 11 participants used voice interfaces daily, 1 used them monthly, and 2 never used them. Among the daily users, 3 used wrist-worn devices (e.g., Apple Watch), and 8 used mobile phones or smart home devices (e.g., Amazon Echo). Voice interfaces were primarily used for simple tasks such as checking the weather, controlling lights, setting alarms, and voice-based texting.

Experience with AI-enabled chatbots: Compared to voice interfaces, participants were less frequent users of AI chatbots with visual interfaces (e.g., ChatGPT), with 7 using them daily, 4 weekly, and 3 once a month or less. Chatbots were primarily used for writing tasks, software development, and exploring new topics for fun.

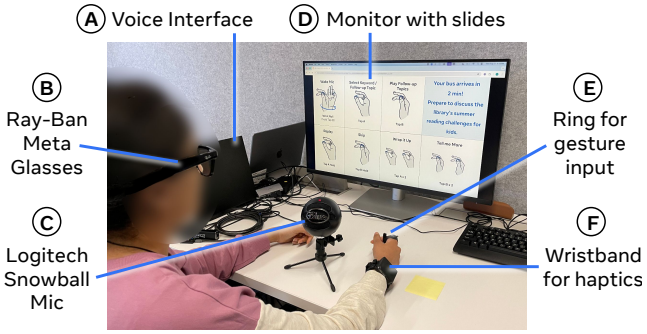

Figure 11: User Study Set-Up. Participants in our exploratory user study were instrumented with Ray-Ban Meta Glasses for audio output *(B)* and in-house wearables for gesture input *(E)* and haptics *(F)*. The LLM-enabled voice interface ran on a laptop *(A)*, connected to a Logitech Snowball microphone *(C)* to capture participants' speech. We displayed task instructions and illustrations of the gesture set on a monitor *(D)*.

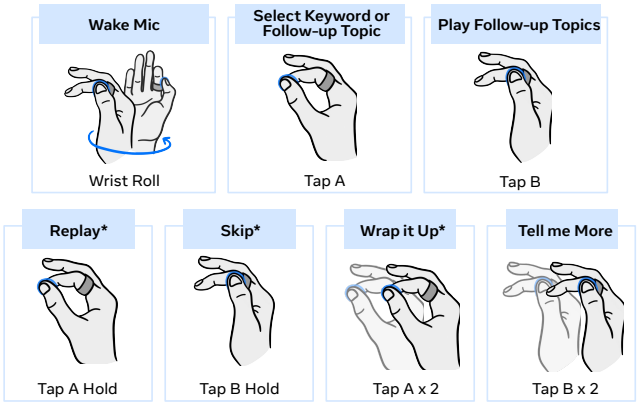

Figure 12: Mapping of Gestures to System Functions. For our user study, we designed a gesture set leveraging three distinct poses (Wrist Roll, Tap A, Tap B) with variations of the tap gestures (holding or repeated taps) to distinguish between system functions. This mapping is a simplified version of the gesture set described in Sec. 3. (*) indicates gestures introduced in Task 2.

5.2 Method

Figure 11 shows the components of our study apparatus. The study consisted of a training session and two scenario-based tasks, where participants conversed with our LLM-enabled voice interface (Sec. 4) using two different interaction styles:

- **Voice-Driven Interaction:** Users specify requests via voice and receive responses via audio. Invoking a *wake word* ("`Hey Roger`") can be used to activate the system and speak a prompt, as well as to interrupt the system's response. We provide audio feedback to indicate when the mic is active and when responses are loading. This traditional interaction style served as a baseline to compare our gesture and haptic techniques against.
- **Gesture+Haptic-Guided Interaction:** In addition to using voice and audio for conversation content, users can control conversation flows using a subset of gesture and haptic guidance techniques from our design space (Fig. 3). Similar to the *wake word* in the Voice-Driven condition, a *wake gesture* (`Wrist Roll`) can be used to initiate a spoken prompt and interrupt the system. The system provides the full range of audio-haptic feedback that we designed to support users' awareness of system state and confirmation of their gestural input. Additionally, the system nudges users to take future actions (e.g., exploring keywords) via haptic-only feedback.

Training: Conversing via Voice and Gesture-Based Interaction Techniques *(15 min)*. We familiarized participants with traditional Voice-Driven interaction through a conversation about their country of residence. Then, using the same conversation topic, we walked through a subset of four Gesture+Haptic-Guided interactions available during the first task (Fig. 12): `Waking the Mic` to pose a prompt, `Playing Follow-up Topics`, `Selecting Keywords or Follow-up Topics` to branch into new directions, and `Tell me More` to keep exploring the current topic.

Conversation Task 1: Learning about an Unfamiliar Topic *(25 min)*. First, we explored the affordances of gesture and haptic guidance techniques for the common LLM use case of *information foraging* [42] (i.e., probing a vast information space to deepen one's understanding of a specific facet). This task was selected for its known challenges with navigational and temporal flow (e.g., parsing long responses and traversing branches of discussion [52]).

Participants chose an unfamiliar topic from a prescribed list (e.g., "how manual clocks work," "uses of coriander"). Starting with either the Voice-Driven or Gesture+Haptic-Guided interaction styles, they conducted a 2-min conversation with the voice interface, aiming to remember details about the topic. To verify participants' attentiveness, we asked them to teach the topic back to the researchers. Following the conversation, we discussed aspects of the system that made it easy or challenging to use.

Participants repeated this task with the interaction style they had not yet experienced, using different conversation topics. We counterbalanced the order of the Voice-Driven and Gesture+Haptic-Guided conditions across participants. For the Gesture+Haptic-Guided condition, participants could control the flow of the conversation using the subset of gestures they learned in the training session (`Waking the Mic`, `Playing Follow-up Topics`, `Selecting Keywords or Follow-up Topics`, and `Tell me More`).

In a semi-structured interview portion after both conversations, we again discussed the easy and challenging aspects of the system usage and participants' preferences for interaction techniques. We also asked participants to compare both conditions and comment on how, if at all, the way they navigated the assistant's responses and managed the timing of the conversation changed.

Conversation Task 2: Preparing for a Meeting *(25 min)*. In contrast with Task 1, which focused on in-*depth* topic exploration, Task 2 required participants to quickly summarize a *breadth* of information. We established a scenario of using a voice assistant to prepare for an upcoming meeting while commuting to work. This scenario was intended to induce time-pressure for participants, enabling us to compare the affordances of voice vs. gestures for

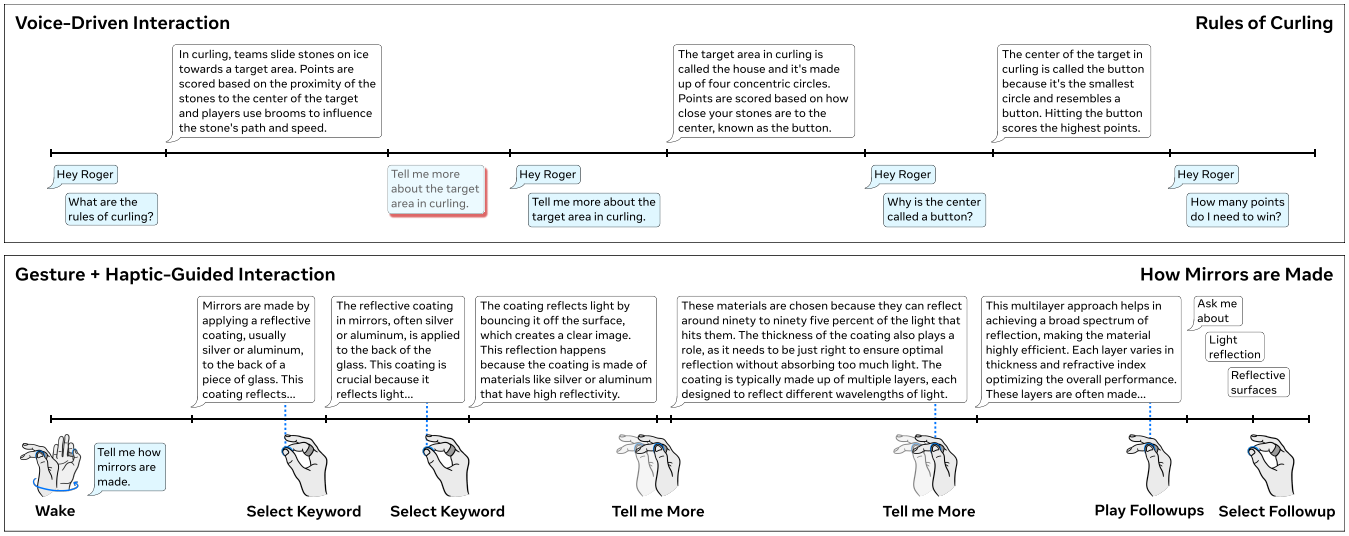

Figure 13: Example Conversation Timelines. We show a partial recreation of P6's Task 1 (Learning) conversations, where they learned about the rules of curling (via Voice-Driven interaction) and how mirrors are made (via Gesture+Haptic-Guided interaction). Light blue speech bubbles denote user speech, while white speech bubbles represent system speech. Interruptive gestures, such as `Select Keyword` during system speech, are depicted as blue dotted lines.

facilitating more active discussions and user-initiated interruptions. In addition to the functions available in Task 1 (Fig. 12), we incorporated three gestures to support faster-paced conversations: `Replay`, `Skip`, and `Wrap It Up` (which finishes reading the current sentence before stopping). New gestures were staggered across Tasks 1 & 2 to ease the process of learning the full gesture set.

For this task, we implemented a `Meeting Summarizer` GPT agent that answers users' queries with respect to two meeting transcripts–one on social media strategy for a new breakfast cereal and the other on a library summer reading program for kids. We generated the transcripts via GPT-4 through iterative prompting and manual editing to enhance realism. As in Task 1, participants engaged in two 2-min conversations using the Voice-Driven and Gesture+Haptic-Guided interaction techniques. To mitigate learning effects, we reversed the order of interaction conditions performed in Task 1.

Following each condition, participants completed the *Subjective Assessment of Speech User Interfaces (SASSI)* questionnaire [17], which holistically assesses user experience across six dimensions: system response accuracy, likeability, cognitive demand, annoyance, habitability, and speed. We used this scale to confirm that the system met usability standards and elicit additional context about the differences between interaction styles. We ended with a discussion around how participants managed navigating and timing within conversations in both conditions.

5.3 Data Collection & Analysis

We captured audio recordings of all study sessions and screen recordings of the study interface. Based on the audio transcripts, we used an affinity diagramming approach [47] to aggregate themes around the easy and challenging aspects of the system and considerations for navigational and temporal flow.

We also recorded timestamped event logs via our Unity client, including instances of user and system speech, gestures, and haptic nudges. This enabled two types of analysis across the Voice-Driven and Gesture+Haptic-Guided conditions: *(1)* a quantitative analysis of voice and gesture inputs, interruptions, and errors; *(2)* a visual comparison of conversation behaviors via timeline visualizations (Fig. 13, 15, Supplemental Material) to confirm our qualitative analysis of participants' feedback. These timelines were created by plotting event logs with the Pyplot library in Python and manually overlaying visuals in Figma.

We used a Wilcoxon signed-rank test for two paired samples to analyze differences across conditions in voice and gesture input frequency, errors, and *SASSI* [17] ratings.

6 Results

Figure 13 shows Participant 6's Task 1 (Learning) conversation timelines for both the Voice-Driven and Gesture+Haptic-Guided conditions. These timelines demonstrate key trends we observed across participants: more balanced turn-taking and slower-paced interaction with Voice-Driven interaction; more efficient accessing of information and interruptive actions to direct conversations with Gesture+Haptic-Guided interaction.

In this section, we first present a quantitative analysis of participants' voice and gesture interactions across tasks and their UX ratings from the *SASSI* [17] questionnaire (Sec. 6.1-6.2). Then, we discuss six qualitative themes highlighting the benefits and challenges that voice- and gesture-driven interaction introduced for navigating and managing timing of conversations with our LLM-enabled voice interface (Sec. 6.3-6.4).

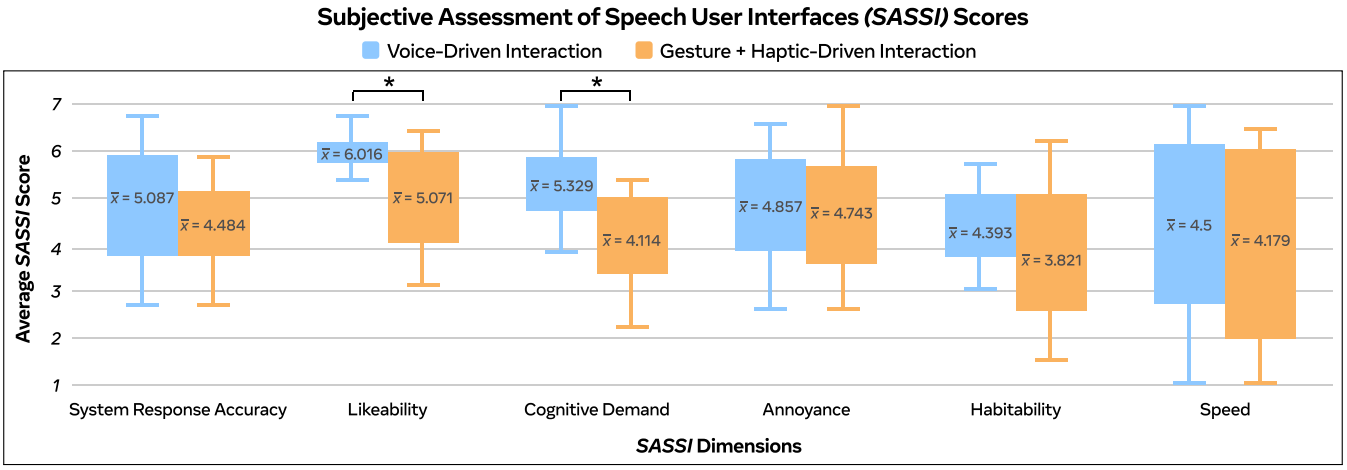

Figure 14: Subjective Assessment of Speech User Interfaces (SASSI) Responses. Box-and-whisker plot showing the range and mean ($\bar{x}$) of SASSI questionnaire ratings across its six UX dimensions. Ratings are on a 7-point Likert scale, where higher scores indicate more favorable responses (1 = Strongly Disagree, 4 = Neutral, 7 = Strongly Agree). Dimensions where the Voice-Driven condition scored significantly higher than the Gesture+Haptic-Guided condition (Likeability and Cognitive Demand) are marked with (*).

6.1 Voice & Gesture Interaction Characteristics

Our analysis revealed significantly more user inputs and interruptions of system speech in the Gesture+Haptic-Guided condition, with no significant difference in error rates between voice and gesture inputs. Full statistical results are provided in Appendix A.1.

Frequency and Type of Interactions. We define a user interaction as the invocation of the *wake gesture* or *wake word* with a corresponding verbal prompt, or other nonverbal gestures. Participants performed significantly more interactions (248) in the Gesture+Haptic-Guided condition compared to 137 in the Voice-Driven condition (Z=-3.296, p=0.001). In Task 1 (Learning), the most popular gestures were `Wake` (35), `Play Follow-up Topics` (19), and `Tell me More` (16). Although we introduced additional gestures to support faster-paced conversations in Task 2 (Meeting Prep), i.e., `Wrap it Up` and `Skip`, the most popular functions were `Wake` (31), `Replay` (20), and `Tell me More` (15).

Interruptive Inputs. We analyzed participants' interruptions of system speech, including use of the *wake word* or gestures that immediately stop or modify system speech (`Wake`, `Select Keyword`, `Play Follow-up Topics`, `Skip`, `Replay`). The number of interruptions was significantly higher in the Gesture+Haptic-Guided condition (58), compared to 8 in the Voice-Driven condition (Z=-3.18, p=0.002). Comparing across tasks, we observed significantly more interruptions in the faster-paced Task 2 (Meeting Prep; *mean*=2.79x per conversation) than in Task 1 (Learning; *mean*=1.36x) for the Gesture+Haptic-Guided condition (Z=-2.04, p=0.043). No significant effects were observed between tasks for Voice-Driven interaction (*mean*=0.21x and 0.36x for Tasks 1 and 2, respectively).

Input Errors. Our analysis identified 28 voice-input errors (*error rate*=0.137 across both the Voice-Driven and Gesture+Haptic-Guided conditions) and 52 gesture-input errors (*error rate*=0.203) across participants. A Wilcoxon signed-rank test found no significant difference in error rates (Z=-1.538, p=0.132). Note that false negatives that were not registered as speech or touch events in our system may not be reflected in these rates.

Voice errors included premature speech processing by the Meta Voice SDK during user pauses (11), incorrect transcription (e.g., "pages" instead of "badges," 10), cut-off speech due to users speaking before the mic activated (4), and forgetting to use the wake word (3).

Gesture errors included performing gesture sequences outside of our gesture set (e.g., triple tap or swiping; 31 instances) and using inactive gestures (19 instances), such as attempting `Select Follow-up` long after all follow-ups were read. Many of the 19 inactive gesture errors involved Tap A and B; based on when participants encountered these errors in the conversation timelines, we infer that 8 errors resulted from participants confusing Tap A with Tap B or vice versa. Additionally, P2 reported 2 instances of unintentionally triggering Tap A gestures while keeping their hands in a resting position; other such instances may have gone unreported by other participants.

6.2 *Subjective Assessment of Speech User Interfaces* Ratings

Figure 14 visualizes the mean and range of participants' *SASSI* ratings for both study conditions, stratified by UX dimension. Appendix A.2 visualizes responses for all 34 questions and provides the full statistical analysis.

Participants rated the Voice-Driven condition significantly higher for Likeability (Z=2.731, p=0.007) and Cognitive Demand (Z=2.605, p=0.01). No significant differences were found for System Response Accuracy, Annoyance, Habitability, or Speed. In post-task discussions, participants attributed the higher Voice-Driven ratings to their familiarity and expertise with voice interfaces, expecting their comfort with gestures to improve as they practiced and memorized the gesture set (P2, P4, P8-11).

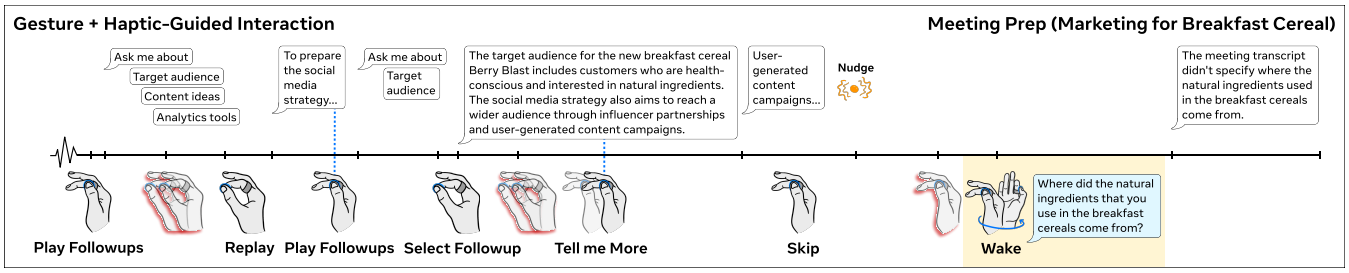

Figure 15: Lack of Multimodal Input. A common "side-effect" of the Gesture+Haptic-Guided interaction techniques was the infrequent use of voice input, which participants could activate via the `Wake` gesture. Here, P1 resists using voice input until the end of their Task 2 (Meeting Prep) conversation, despite performing a few unintended gestures that derailed the conversation.

6.3 Navigational Flow

Next, we discuss three qualitative themes around how participants leveraged voice and gestures to navigate the system's responses. Participants noted a difference in who was leading the conversation: Voice-Driven interaction felt more user-driven, while Gesture+Haptic-Guided interaction felt more system-driven. Turn-taking behaviors also varied across conditions: Voice-Driven interaction facilitated a balanced cadence of user and system speech, while gestures supported more interruptive actions.

Voice-only conversations were perceived as user-driven with open-ended possible actions. Based on the *SASSI* questionnaire responses (Appendix A.2), all participants felt in control of the conversation when using Voice-Driven interaction. A majority of participants attributed this sense of leadership to their familiarity with the system's interaction language (P2, P4-12), as they "don't have to learn how to speak" (P2), and the expressiveness of the voice modality, as they "can ask for whatever [they] want" (P12).

However, participants experienced a loss of control when they were "stumped on what to ask" (P7) or lacked a mental model of the floor and ceiling of the system's capabilities, in line with prior empirical studies of voice interfaces [32]. P3 expressed that voice-only interaction was "too flexible," as they "could end up going down a tangent" without an "indication of what's important."

With gesture & haptic guidance techniques, conversations were perceived as system-driven with constrained possible actions. Participants found that Gesture+Haptic-Guided interaction introduced an extra step of translating their own intent into specific commands in the system's gesture language (P1, P3-5, P8-9, P11-12, P14). As P8 explained, "I was not thinking about *what* I want to ask; I was more thinking [about] *how* I can ask it using gestures." The scoped set of gesture-based functions, along with the system's haptic nudges to highlight possible follow-up actions, gave participants the impression that "the leader in the conversation is the voice assistant" (P5).

These system-driven operations benefited participants by "adding structure" to the conversation (P9) and demonstrating effective ways to prompt the system using voice: "knowing that the assistant can add a follow up helped me realize what all I can ask" (P2). Several participants adapted strategies from our gesture-based functions to support their conversations in the Voice-Driven condition (P2, P8, P11, P13-14), e.g., speaking "`Tell me More`" and "`Replay.`"

However, participants tended to fixate on the constrained set of gestures rather than verbally clarifying their intent. On average, only 28.36% of participants' gestures were used to `Wake` the mic, compared to other nonverbal commands. We observed this adherence to gestures even when conversations went off track. For example, P1 struggled to obtain a useful summary in Task 2 (Meeting Prep) due to consecutively performing incorrect gestures but did not switch to voice input until the end of the 2-min period (Fig. 15). Participants were hesitant to use multimodal interaction for various reasons: preferring subtle gestures over voice input in a public context for the Task 2 (Meeting Prep) scenario (P3), being more comfortable performing taps than the `WristRoll` gesture to activate the mic (P7), and lacking a mental model for which interactions are best suited for voice versus gestures (P8).

Voice-only interaction led to balanced turn-taking and complete exchanges; gestures facilitated more interruptive actions to direct conversations. The Voice-Driven condition resulted in significantly fewer interruptions than gesture-driven interaction (Sec. 6.1); participants typically allowed the system to finish speaking before posing new prompts, even when it delivered repetitive or irrelevant information. Participants attributed this hesitance to social norms in human-to-human interaction (P2, P4-5, P9, P13). P4 expressed, "it felt like I'm talking to somebody... so I let him finish the sentence, because it felt rude to just stop." In contrast, interrupting with gestures was considered more socially acceptable: "I don't have to commit the act of interrupting by voice. I just kind of send a signal" to the system (P13).

6.4 Temporal Flow

Finally, we present three themes around how participants managed the timing of conversations in either condition. They found gestures to support more efficient information access, but could more quickly verify the system's alignment with their intent when using voice-driven interaction. Additionally, haptic cues interleaved with system speech increased participants' perception of time pressure.

Accessing information with gesture & haptic-guided interaction required less perceived time and effort. Participants expressed that the combination of gestures and audio-haptic cues increased their awareness of possible ways to extract details from the voice interface, like "having information at your fingertips" (P7).

With Voice-Driven interaction, many participants noted that speaking the Wake Phrase disrupted the flow of conversation (P1, P3, P6-7, P10-P13) and found it "bothersome and slow" to "ask all questions [themselves]" (P6). In contrast, participants felt gestures offered more expressive power [20], often yielding a similar amount and quality of information while requiring less effort (P2, P6-7, P9, P11, P13-14): "I can ask [the system] to elaborate on what it knows instead of having to think about what I should ask it" (P2). This difference between conditions is also evident from our timeline analysis: conversations in the Voice-Driven condition were sparser, whereas the Gesture+Haptic-Guided conversations had more instances of system speech surfacing new details (Fig. 13).

Direct specification of prompts through voice saves time in the gulf of evaluation; indirect specification through gestures can cause uncertainty about the system's understanding. When using voice to explicitly pose prompts to the voice interface, some participants felt "better aware of the response" (P5) and had more confidence that the system understood "what type of interaction [they] wanted" (P1). However, when guiding the conversation via gestures, they sometimes lacked a clear understanding of how the system was acting upon their input (P1, P4, P5, P7-8, P14). This uncertainty primarily arose for the `Select Keyword` interaction, which required precisely timing gestures to target words in the system's response and waiting a few seconds to receive a new response. P5 expressed: "I feel distracted because I don't have full trust that my response is going to be interpreted in a timely manner... which is taking up space from me retaining information."

Voice-only interaction led to slower paced conversations; gesture & haptic-guided interaction induced more time pressure. As discussed in Sec. 6.3, participants felt it socially unacceptable to interrupt in the Voice-Driven condition, which led to a more leisurely turn-taking pace. Participants appreciated the slower pace for enabling them to carefully listen to and process information in Task 1 (Learning; P3-6, P8-9, P13), but some considered their performance in Task 2 (Meeting Prep) to be more successful in the Gesture+Haptic-Guided condition, as gestures better lent themselves to efficiently summarizing topics (P2, P5, P7, P9, P13).

A side effect of the system's haptic nudges on keywords, combined with the ease and speed of gestures, was an increased sense of time pressure (P3, P5-6, P8-9, P12-13). While some participants felt that "haptics kept [them] interested" in the conversation (P7), others found their attention divided, as if "multi-tasking between hearing and feeling" (P13). Despite being able to replay responses or verbally specify topics to dive deeper into, participants felt pressured to act immediately on haptic nudges to avoid missed opportunities. P6 described this as being "in a chase" with the system, wanting "more time and space to think" about their next actions. While participants' *SASSI* ratings for the Speed dimension showed no significant differences, most reported that Gesture+Haptic-Guided interaction required high concentration and made them feel tense.

7 Discussion

With the broader goal of enabling users to effectively leverage intelligent capabilities of wearable voice interfaces, we developed a suite of interaction techniques that integrate gestures and haptics as parallel I/O channels alongside voice and audio, building on prior HCI research in conversational and multimodal interaction [22, 59].

Our evaluation of these techniques demonstrated that gestures and haptics afforded participants more flexible control over the navigational and temporal flow of conversations. Specifically, they enabled *(1)* more efficient information access, *(2)* socially acceptable techniques for interrupting and re-directing system speech to align with users' goals, and *(3)* greater awareness of *what* future actions are possible and *how* to prompt the voice interface to execute them.

However, our interaction techniques also introduced unintended side effects, reflected by participants' feedback and *SASSI* ratings: *(1)* hesitation to use multimodal input, *(2)* uncertainty about the system's understanding of intent when using gestures to indirectly prompt the system, and *(3)* a sense of time pressure caused by haptics interleaved with system speech. Participants attributed some of these challenges to the strict roles assigned to I/O modalities (i.e., voice and audio for conversation *content* and gestures and haptics for *control*). While this separation of roles was necessary to enable efficient multi-tasking, there is a need for flexibility.

We propose three design recommendations to improve the integration of these multimodal I/O techniques. In particular, we identify opportunities for voice to complement gestures in clarifying users' intent and for verbal audio to work alongside audio-haptic cues to make the system state more transparent.

Facilitate automatic transitions between gestures and voice input. Participants avoided using voice when gestures were available, as they perceived transitioning between input techniques as high effort. To promote more natural voice-driven communication, we envision two approaches. First, the microphone could automatically activate after a gesture, allowing users to clarify their intent with brief voice commands without needing to manually `Wake` the system. Second, the output channels could be leveraged to guide users toward the most suitable input technique for a given system function (linking audio to voice and haptics to gestures). In our current implementation, haptic feedback during system speech indicates opportunities to silently interrupt or redirect the conversation via gestures. Extending this, audio feedback while the system loads could prompt users to refine their request via voice.

Confirm not only *that* the system received users' input, but also *how* it interpreted users' intent. Gesture-driven interactions that implicitly convey user intent (e.g., `Tell Me More` or `Go Deeper`) led participants to doubt whether the system fully understood them. This suggests that the audio-haptic feedback confirming gestural input was not always descriptive enough to communicate the system's state. To provide more transparency, participants recommended incorporating verbal audio feedback to recap the contextual details detected by the system. For example, during a `Go Deeper` interaction, the system could repeat the keyword it believes the user is focusing on, allowing users to quickly correct misunderstandings.

Offer advanced notice and delayed actions to reduce time pressure. Our haptic nudges interleaved with system speech to highlight interaction opportunities led some participants to fixate on the nudges and feel compelled to act immediately. To alleviate time pressure, we propose two strategies. First, haptic patterns could be adjusted from short bursts to a series of pulses to give users'

advanced notice that an opportunity is approaching, e.g., "ramping up" to a keyword event or signaling the end of system speech. Second, gesture variations could distinguish between immediate and delayed actions (e.g., `Pinch` could translate to `Go Deeper Now`, while `Pinch Hold` could queue up detailed responses for later).

8 Limitations

We discuss the main limitations of our research: generalizability of our system and study to other form factors and usage scenarios, the need to study with novice and differently-abled users, and potential novelty effects.

Generalizability to other form factors and sensing approaches: We used a ring-based gesture input device that supports a subtle interaction style (i.e., on-skin gestures that users can perform with hands on their lap or to their side). Future studies should explore to what extent our findings generalize to other gesture recognition approaches. In particular, we anticipate that users' perception of subtle gestures as a socially acceptable alternative to voice interruptions may not hold for camera-based gesture recognition, which future smart glasses could enable.

Generalizability to in-the-wild usage scenarios: Our studies were conducted in a controlled lab environment; however gesture and speech recognition performance would vary in the wild. For example, noisy environments increase the risks of false positives and negatives in speech recognition, while users' interactions with real-world objects can lead to false positives in gesture recognition. Future work could explore alternative sensing techniques or gesture sets to accommodate multi-tasking (e.g., microgestures that transfer across different hand poses and location constraints [19]).

Study sample: A majority of study participants were experienced voice interface users; as such, their insights into the benefits and challenges of Voice-Driven and Gesture+Haptic-Guided interaction may not generalize to novice users. However, our results suggest that our gesture-driven interaction techniques are approachable for novices, given that all participants learned the gestures within a 10-min training period and successfully completed the study tasks. Further research is required to explore the needs of users with hearing or motor impairments. Personalization mechanisms could further improve the accessibility of the interaction techniques (e.g., adjusting voice speed and choosing custom gesture mappings that are more memorable and comfortable).

Novelty effects: Finally, given the novelty of the gesture and haptic-driven interactions, our results could be subject to participant response bias [11]. We mitigated potential bias by probing into both the easy and challenging aspects of the interaction techniques. Participants' *SASSI* questionnaire responses rated the Voice-Driven condition more favorably for Likeability and Cognitive Demand, with similar ratings across other UX dimensions. This suggests that participants did not exhibit a bias toward the Gesture+Haptic-Guided condition.

9 Future Work & Conclusion

To reduce friction in conducting complex conversations with intelligent voice interfaces, we developed a suite of multimodal interaction techniques that introduce gestures and haptics as alternate I/O modalities to voice and audio. This enables a role-based collaboration: voice and audio focus on conversation content, gestures control conversation flow, and haptics convey interaction opportunities. Our development was informed by a review and classification of multimodal interaction in existing conversational UIs. Through a study with 14 end-users, we found that our gesture and haptic guidance techniques enabled more efficient information access, offered socially acceptable alternatives to interrupting system speech, and raised participants' awareness of future actions they could take in the conversation. To address challenges faced by participants, we propose design recommendations to smoothly transition between I/O modalities and reduce time pressure when interleaving gesture-based interactions with system speech.

We identify two promising avenues for future work. First, technical improvements could explore mixed-initiative approaches to interpreting the meaning of gestures to minimize users' effort. For example, a one-to-many mapping of gestures to system functions could enable a smaller, more memorable gesture set by inferring meaning from the timing and context of the conversation. Users' speech and gesture characteristics could be leveraged in implicit interactions to guide conversation flows, e.g., users inhaling or repositioning their hands could signal a desire to speak, prompting the system to `Wrap Up` its response. To reduce the need for users to prompt with follow-up actions, the system could automatically resume the main conversation thread after a `Go Deeper` interaction.

Second, future empirical studies could investigate considerations for using similar multimodal interaction techniques in real-world contexts. As next generation smartglasses increasingly integrate camera-based input, it would be interesting to explore how our techniques translate to new types of conversations such as world-querying, lifelogging, or receiving guidance on physical tasks.

Acknowledgments

We thank Daylon Walden for designing our voice interfaces' audio-haptic cues and Roger Boldu & Dan Clarke for their advice on the gesture sensing pipeline. We also thank Tovi Grossman, Dana Sasinowski, and our other colleagues at Reality Labs Research for their feedback and support throughout the project.

References

[1] Matthew Peter Aylett and Marta Romeo. 2023. You don't need to speak, you need to liste: Robot interaction and human-like turn-taking. In *Proc. of the 5th Int. Conf. on Conversational User Interfaces*. 1–5.

[2] Jose Daniel Azofeifa, Julieta Noguez, Sergio Ruiz, José Martín Molina-Espinosa, Alejandra J Magana, and Bedrich Benes. 2022. Systematic review of multimodal human-computer interaction. In *Informatics*, Vol. 9. MDPI, 13.

[3] Alan Baddeley. 1992. Working memory. *Science* 255, 5044 (1992), 556–559. doi:10.1126/science.1736359

[4] Kristoffer Bergram, Marija Djokovic, Valéry Bezençon, and Adrian Holzer. 2022. The digital landscape of nudging: A systematic literature review of empirical research on digital nudges. In *Proc. ACM Conf. on Human Factors in Computing Systems*. ACM, 62:1–62:16. doi:10.1145/3491102.3517638

[5] Valeria Bolotova, Vladislav Blinov, Falk Scholer, W. Bruce Croft, and Mark Sanderson. 2022. A non-factoid question-answering taxonomy. In *Proc. ACM Int. Conference on Research and Development in Information Retrieval (SIGIR)*. ACM, 1196–1207. doi:10.1145/3477495.3531926

[6] Richard A. Bolt. 1980. "Put-that-there": Voice and gesture at the graphics interface. In *Proc. Conf. on Computer Graphics and Interactive Techniques (SIGGRAPH)*. ACM, 262–270. doi:10.1145/800250.807503

[7] Runze Cai, Nuwan Janaka, Yang Chen, Lucia J. Wang, Shengdong Zhao, and Can Liu. 2024. PANDALens: Towards AI-assisted in-context writing on OHMD

during travels. In *Proc. ACM Conf. on Human Factors in Computing Systems*. ACM, 1053:1–1053:24. doi:10.1145/3613904.3642320

[8] P. A. Chandler and J. Sweller. 1992. The split-attention effect as a factor in the design of instruction. *British Journal of Educational Psychology* 62 (1992), 233–246.

[9] Neila Chettaoui, Ayman Atia, and Med Salim Bouhlel. 2022. Exploring the impact of interaction modality on students' learning performance. *J. Educational Computing Research* 60, 1 (2022), 4–27.

[10] Philip R. Cohen, Michael Johnston, David McGee, Sharon L. Oviatt, Jay Pittman, Ira A. Smith, Liang Chen, and Josh Clow. 1997. QuickSet: Multimodal interaction for distributed applications. In *Proc. of the ACM Int. Conf. on Multimedia*. ACM Press, 31–40. doi:10.1145/266180.266328

[11] Nicola Dell, Vidya Vaidyanathan, Indrani Medhi, Edward Cutrell, and William Thies. 2012. "Yours is better!": Participant response bias in HCI. In *Proc. ACM Conf. on Human Factors in Computing Systems*. ACM, 1321–1330. doi:10.1145/2207676.2208589

[12] Bruno Dumas, Denis Lalanne, and Sharon L. Oviatt. 2009. Multimodal interfaces: A survey of principles, models and frameworks. In *Human Machine Interaction*. https://api.semanticscholar.org/CorpusID:16540884

[13] Ryan Khushan Ghamandi, Ravi Kiran Kattoju, Yahya Hmaiti, Mykola Maslych, Eugene Matthew Taranta, Ryan P. McMahan, and Joseph J. LaViola. 2024. Unlocking understanding: An investigation of multimodal communication in virtual reality collaboration. In *Proc. ACM Conf. on Human Factors in Computing Systems*. ACM, 602:1–602:16. doi:10.1145/3613904.3642491

[14] Sanjay Ghosh, Anirudha Joshi, and Sanjay Tripathi. 2013. Empirical evaluation of multimodal input interactions. In *Human Interface and the Management of Information. Information and Interaction Design*, Sakae Yamamoto (Ed.). Springer, Berlin, Heidelberg, 37–47.

[15] Gabriel Haas, Michael Rietzler, Matt Jones, and Enrico Rukzio. 2022. Keep it short: A comparison of voice assistant' response behavior. In *Proc. ACM Conf. on Human Factors in Computing Systems*. ACM, 321:1–321:12. doi:10.1145/3491102.3517684

[16] Ken Hinckley, Koji Yatani, Michel Pahud, Nicole Coddington, Jenny Rodenhouse, Andy D. Wilson, Hrvoje Benko, and Bill Buxton. 2010. Pen + touch = new tools. In *Proc. ACM Symposium on User Interface Software and Technology*. ACM, 27–36. doi:10.1145/1866029.1866036

[17] Kate Hone, Ub Ph, Robert Graham, and Alencon Link. 2000. Towards a tool for the subjective assessment of speech system interfaces (SASSI). *Natural Language Engineering* 6 (07 2000). doi:10.1017/S1351324900002497

[18] Peiling Jiang, Jude Rayan, Steven P. Dow, and Haijun Xia. 2023. Graphologue: Exploring large language model responses with interactive diagrams. In *Proc. ACM Symposium on User Interface Software and Technology*. ACM, 3:1–3:20. doi:10.1145/3586183.3606737

[19] Nikhita Joshi, Parastoo Abtahi, Raj Sodhi, Nitzan Bartov, Jackson Rushing, Christopher Collins, Daniel Vogel, and Michael Glueck. 2023. Transferable microgestures across hand posture and location constraints: Leveraging the middle, ring, and pinky fingers. In *Proc. ACM Symposium on User Interface Software and Technology*. ACM, 103:1–103:17. doi:10.1145/3586183.3606713

[20] Dan R. Olsen Jr. 2007. Evaluating user interface systems research. In *Proc. ACM Symposium on User Interface Software and Technology*, Chia Shen, Robert J. K. Jacob, and Ravin Balakrishnan (Eds.). ACM, 251–258. doi:10.1145/1294211.1294256

[21] Jin Gun Jung, Sangyoon Lee, Jiwoo Hong, Eunhye Youn, and Geehyuk Lee. 2020. Voice+Tactile: Augmenting in-vehicle voicer user interface with tactile touchpad interaction. In *Proc. ACM Conf. on Human Factors in Computing Systems*. ACM, 1–12. doi:10.1145/3313831.3376863

[22] Runchang Kang, Anhong Guo, Gierad Laput, Yang Li, and Xiang 'Anthony' Chen. 2019. Minuet: Multimodal interaction with an Internet of Things. In *Proc. Symp. on Spatial User Interaction, SUI*. ACM, 2:1–2:10. doi:10.1145/3357251.3357581

[23] Wolf Kienzle, Eric Whitmire, Chris Rittaler, and Hrvoje Benko. 2021. Electroring: Subtle pinch and touch detection with a ring. In *Proc. ACM Conf. on Human Factors in Computing Systems*. 1–12.

[24] Jung Kim, Hyun Kim, Boon K. Tay, Manivannan Muniyandi, Joel Jordan, Jesper Mortensen, Manuel Oliveira, Mel Slater, and Mandayam A. Srinivasan. 2004. Transatlantic Touch: A study of haptic collaboration over long distances. *Presence Teleoperators Virtual Environ.* 13, 3 (2004), 328–337. doi:10.1162/1054746041422370

[25] Tae Soo Kim, Arghya Sarkar, Yoonjoo Lee, Minsuk Chang, and Juho Kim. 2023. LMCanvas: Object-oriented interaction to personalize large language model-powered writing environments. arXiv:2303.15125 [cs.HC] https://arxiv.org/abs/2303.15125

[26] Yoonsu Kim, Jueon Lee, Seoyoung Kim, Jaehyuk Park, and Juho Kim. 2024. Understanding users' dissatisfaction with ChatGPT responses: Types, resolving tactics, and the effect of knowledge leve. In *Proc. of the ACM Int. Conf. on Intelligent User Interfaces (IUI)*. ACM, 385–404. doi:10.1145/3640543.3645148

[27] Divesh Lala, Koji Inoue, and Tatsuya Kawahara. 2019. Smooth turn-taking by a robot using an online continuous model to generate turn-taking cues. In *Proc. Int. Conf on Multimodal Interaction, ICMI*, Wen Gao, Helen Mei-Ling Meng, Matthew A. Turk, Susan R. Fussell, Björn W. Schuller, Yale Song, and Kai Yu (Eds.). ACM, 226–234. doi:10.1145/3340555.3353727

[28] Raina Langevin, Ross J. Lordon, Thi Avrahami, Benjamin R. Cowan, Tad Hirsch, and Gary Hsieh. 2021. Heuristic evaluation of conversational agents. In *Proc. ACM Conf. on Human Factors in Computing Systems*. ACM, 632:1–632:15. doi:10.1145/3411764.3445312

[29] Jaewook Lee, Jun Wang, Elizabeth Brown, Liam Chu, Sebastian S. Rodriguez, and Jon E. Froehlich. 2024. GazePointAR: A context-aware multimodal voice assistant for pronoun disambiguation in wearable augmented reality. In *Proc. ACM Conf. on Human Factors in Computing Systems*. ACM, 408:1–408:20. doi:10.1145/3613904.3642230

[30] John Lee, Christopher Wickens, Yili Liu, and Linda Boyle. 2017. *Designing for People: An Introduction to Human Factors Engineering*. CreateSpace. 159–161 pages.

[31] Jiahao Nick Li, Yan Xu, Tovi Grossman, Stephanie Santosa, and Michelle Li. 2024. OmniActions: Predicting digital actions in response to real-world multimodal sensory inputs with LLMs. In *Proc. ACM Conf. on Human Factors in Computing Systems*. ACM, 8:1–8:22. doi:10.1145/3613904.3642068

[32] Ewa Luger and Abigail Sellen. 2016. "Like having a really bad PA": The gulf between user expectation and experience of conversational agents. In *Proc. ACM Conf. on Human Factors in Computing Systems*. ACM, 5286–5297. doi:10.1145/2858036.2858288

[33] Damien Masson, Sylvain Malacria, Géry Casiez, and Daniel Vogel. 2024. DirectGPT: A direct manipulation interface to interact with large language models. In *Proc. ACM Conf. on Human Factors in Computing Systems*. ACM, 975:1–975:16. doi:10.1145/3613904.3642462

[34] Miguel A. Nacenta, Yemliha Kamber, Yizhou Qiang, and Per Ola Kristensson. 2013. Memorability of pre-designed and user-defined gesture sets. In *Proc. ACM Conf. on Human Factors in Computing Systems*. ACM, 1099–1108. doi:10.1145/2470654.2466142

[35] Satoshi Nakamura, Takeshi Shoji, Masahiko Tsukamoto, and Shojiro Nishio. 2005. SoundWeb: Hyperlinked voice data for wearable computing environment. In *Proc. IEEE Int. Symp. on Wearable Computers*. IEEE Computer Society, 14–19. doi:10.1109/ISWC.2005.47

[36] Anja B Naumann, Ina Wechsung, and Jörn Hurtienne. 2010. Multimodal interaction: A suitable strategy for including older users? *Interacting with Computers* 22, 6 (2010), 465–474.

[37] Michael Nebeling and Anind K. Dey. 2016. XDBrowser: User-defined cross-device web page designs. In *Proc. ACM Conf. on Human Factors in Computing Systems*. ACM, 5494–5505. doi:10.1145/2858036.2858048

[38] Fabian Okeke, Michael Sobolev, Nicola Dell, and Deborah Estrin. 2018. Good vibrations: Can a digital nudge reduce digital overload?. In *Proc. ACM Conf. on Human Factors in Computing Systems*. ACM, 4:1–4:12. doi:10.1145/3229434.3229463

[39] Sharon Oviatt. 2003. User-centered modeling and evaluation of multimodal interfaces. *Proc. of the IEEE* 91, 9 (2003), 1457–1468.

[40] Sharon L. Oviatt. 1999. Ten myths of multimodal interaction. *Commun. ACM* 42, 11 (1999), 74–81. doi:10.1145/319382.319398

[41] Antti Pirhonen, Stephen A. Brewster, and Christopher Holguin. 2002. Gestural and audio metaphors as a means of control for mobile devices. In *Proc. ACM Conf. on Human Factors in Computing Systems*, Dennis R. Wixon (Ed.). ACM, 291–298. doi:10.1145/503376.503428

[42] Peter Pirolli and Stuart Card. 1999. Information Foraging. *Psychological Review* 106 (10 1999), 643–675. doi:10.1037/0033-295X.106.4.643

[43] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice interfaces in everyday life. In *Proc. ACM Conf. on Human Factors in Computing Systems*. ACM, 640. doi:10.1145/3173574.3174214

[44] Jeff Raskin. 2000. *The humane interface: New directions for designing interactive systems*. Addison-Wesley.

[45] Nitin "Nick" Sawhney and Chris Schmandt. 1999. Nomadic Radio: Scaleable and contextual notification for wearable audio messaging. In *Proc. ACM Conf. on Human Factors in Computing Systems*, Marian G. Williams and Mark W. Altom (Eds.). ACM, 96–103. doi:10.1145/302979.303005

[46] Chris Schmandt. 1998. Audio Hallway: A virtual acoustic environment for browsing. In *Proc. ACM Symposium on User Interface Software and Technology*, Elizabeth D. Mynatt and Robert J. K. Jacob (Eds.). ACM, 163–170. doi:10.1145/288392.288597

[47] Raymond Scupin. 1997. The KJ Method: A technique for analyzing data derived from Japanese ethnology. *Human Organization* 56, 2 (1997), 233–237.

[48] Abigail J Sellen, Gordon P Kurtenbach, and William AS Buxton. 1992. The prevention of mode errors through sensory feedback. *Human-Computer Interaction* 7, 2 (1992), 141–164.

[49] Vidya Setlur and Melanie Tory. 2022. How do you converse with an analytical chatbot? Revisiting Gricean Maxims for designing analytical conversational behavior. In *Proc. ACM Conf. on Human Factors in Computing Systems*. ACM, 29:1–29:17. doi:10.1145/3491102.3501972

[50] R. Sharma, V.I. Pavlovic, and T.S. Huang. 1998. Toward multimodal human-computer interface. *Proc. of the IEEE* 86, 5 (1998), 853–869. doi:10.1109/5.664275

[51] Sangho Suh, Meng Chen, Bryan Min, Toby Jia-Jun Li, and Haijun Xia. 2024. Luminate: Structured generation and exploration of design space with large

language models for human-AI co-creation. In *Proc. ACM Conf. on Human Factors in Computing Systems*. ACM, 644:1–644:26. doi:10.1145/3613904.3642400

[52] Sangho Suh, Bryan Min, Srishti Palani, and Haijun Xia. 2023. Sensecape: Enabling multilevel exploration and sensemaking with large language models. In *Proc. ACM Symposium on User Interface Software and Technology*. ACM, 1:1–1:18. doi:10.1145/3586183.3606756

[53] Hemant Bhaskar Surale, Aakar Gupta, Mark Hancock, and Daniel Vogel. 2019. TabletInVR: Exploring the design space for using a multi-touch tablet in virtual reality. In *Proc. ACM Conf. on Human Factors in Computing Systems*. ACM, 13. doi:10.1145/3290605.3300243

[54] Hong Z. Tan, Seungmoon Choi, Frances W. Y. Lau, and Freddy Abnousi. 2020. Methodology for maximizing information transmission of haptic devices: A survey. *Proc. of the IEEE* 108, 6 (2020), 945–965. doi:10.1109/JPROC.2020.2992561

[55] Michael Terry, Chinmay Kulkarni, Martin Wattenberg, Lucas Dixon, and Meredith Ringel Morris. 2023. AI alignment in the design of interactive AI: Specification alignment, process alignment, and evaluation support. *CoRR* abs/2311.00710 (2023). doi:10.48550/ARXIV.2311.00710 arXiv:2311.00710

[56] Alexandra Vtyurina, Adam Fourney, Meredith Ringel Morris, Leah Findlater, and Ryen W. White. 2019. VERSE: Bridging screen readers and voice assistants for enhanced eyes-free web search. In *Proc. ACM SIGACCESS Conf. on Computers and Accessibility ASSETS*. ACM, 414–426. doi:10.1145/3308561.3353773

[57] Nina Zhuxiaona Wei and James A. Landay. 2018. Evaluating speech-based smart devices using new usability heuristics. *IEEE Pervasive Comput.* 17, 2 (2018), 84–96. doi:10.1109/MPRV.2018.022511249

[58] Mirjam Wester, David A Braude, Blaise Potard, Matthew P Aylett, and Francesca Shaw. 2017. Real-Time reactive speech synthesis: Incorporating interruptions. In *INTERSPEECH*. 3996–4000.

[59] Liwei Wu, Ben Lafreniere, Tovi Grossman, Thomas White, and Stephanie Santosa. 2024. Body language for VUIs: Exploring gestures to enhance interactions with voice user interfaces. In *Proc. ACM Conf. on Designing Interactive Systems*. ACM. doi:10.1145/3643834.3660691

[60] Shan Xu, Sarah Sykes, Parastoo Abtahi, Tovi Grossman, Daylon Walden, Michael Glueck, and Carine Rognon. 2024. Designing haptic feedback for sequential gestural inputs. In *Proc. ACM Conf. on Human Factors in Computing Systems*. ACM, 711:1–711:17. doi:10.1145/3613904.3642735

[61] Yang Zhang, Wolf Kienzle, Yanjun Ma, Shiu S. Ng, Hrvoje Benko, and Chris Harrison. 2019. ActiTouch: Robust touch detection for on-skinAR/VR interfaces. In *Proc. ACM Symposium on User Interface Software and Technology*. ACM, 1151–1159. doi:10.1145/3332165.3347869

[62] Yang Zhang, Junhan Zhou, Gierad Laput, and Chris Harrison. 2016. SkinTrack: Using the body as an elecrical waveguide for continuous finger tracking on the skin. In *Proc. ACM Conf. on Human Factors in Computing Systems*. ACM, 1491–1503. doi:10.1145/2858036.2858082

[63] Shengdong Zhao, Pierre Dragicevic, Mark H. Chignell, Ravin Balakrishnan, and Patrick Baudisch. 2007. Earpod: eyes-free menu selection using touch input and reactive audio feedback. In *Proc. ACM Conf. on Human Factors in Computing Systems*, Mary Beth Rosson and David J. Gilmore (Eds.). ACM, 1395–1404. doi:10.1145/1240624.1240836

[64] Chen Zhou, Zihan Yan, Ashwin Ram, Yue Gu, Yan Xiang, Can Liu, Yun Huang, Wei Tsang Ooi, and Shengdong Zhao. 2024. GlassMail: Towards personalised wearable assistant for on-the-go email creation on smart glasses. In *Proc. ACM Conf. on Designing Interactive Systems*. ACM. doi:10.1145/3643834.3660683

[65] Fengyuan Zhu and Tovi Grossman. 2020. BISHARE: Exploring bidirectional interactions between smartphones and head-mounted augmented reality. In *Proc. ACM Conf. on Human Factors in Computing Systems*. ACM, 1–14. doi:10.1145/3313831.3376233

A Appendix

A.1 Results: Voice & Gesture Input Characteristics

	Mean	Standard Deviation	Minimum	Maximum
Total Interactions (Voice-Driven)	9.796	1.424	8	12
Total Interactions (Gesture+Haptic-Guided)	17.857	3.92	11	27
Interruptive Inputs (Task 1 Learning, Voice-Driven)	0.214	0.579	0	2
Interruptive Inputs (Task 2 Meeting Prep, Gesture+Haptic-Guided)	0.357	0.497	0	1
Interruptive Inputs (Task 1 Learning, Voice-Driven)	1.357	1.008	0	3
Interruptive Inputs (Task 2 Meeting Prep, Gesture+Haptic-Guided)	2.786	1.626	0	6
Error Rate (Voice-Driven)	0.137	0.111	0.0	0.4
Error Rate (Gesture+Haptic-Guided)	0.203	0.14	0.056	0.444

Table 1: Descriptive Statistics for Voice & Gesture Interaction (across all 14 participants).

	Test Statistic (Z)	p-value	Effect Size (r)
***Total Interactions**	-3.296	**0.001**	-0.881
***Interruptive Inputs** (Task 1 Learning, both conditions)	-2.578	**0.009**	-0.689
***Interruptive Inputs** (Task 2 Meeting Prep, both conditions)	-3.059	**0.002**	-0.818
Interruptive Inputs (Voice-Driven condition, both tasks)	-0.913	0.424	-0.244
***Interruptive Inputs** (Gesture+Haptic-Guided condition, both tasks)	-2.04	**0.043**	-0.545
Error Rate	-1.538	0.132	-0.411

Table 2: Statistical Analysis for Voice & Gesture Interaction Characteristics. The data was analyzed via a Wilcoxon signed-rank test for two paired samples; significant differences are indicated by (*). Due to the small sample size, we also report the effect size to assess the magnitude of observed differences. For all statistically significant differences ($p < 0.05$), we observed a large effect size ($|r| > 0.5$).

We observed a significantly higher number of total interactions and interruptive inputs for both tasks in the Gesture+Haptic-Guided condition. Comparing across tasks, we observed significantly more interruptions in the faster-paced Task 2 (Meeting Prep) than in Task 1 (Learning) for the Gesture+Haptic-Guided condition. No significant effects were observed between tasks for Voice-Driven interaction. Additionally, no significant differences between error rate for voice and gesture input were observed.

A.2 Results: Subjective Assessment of Speech User Interfaces

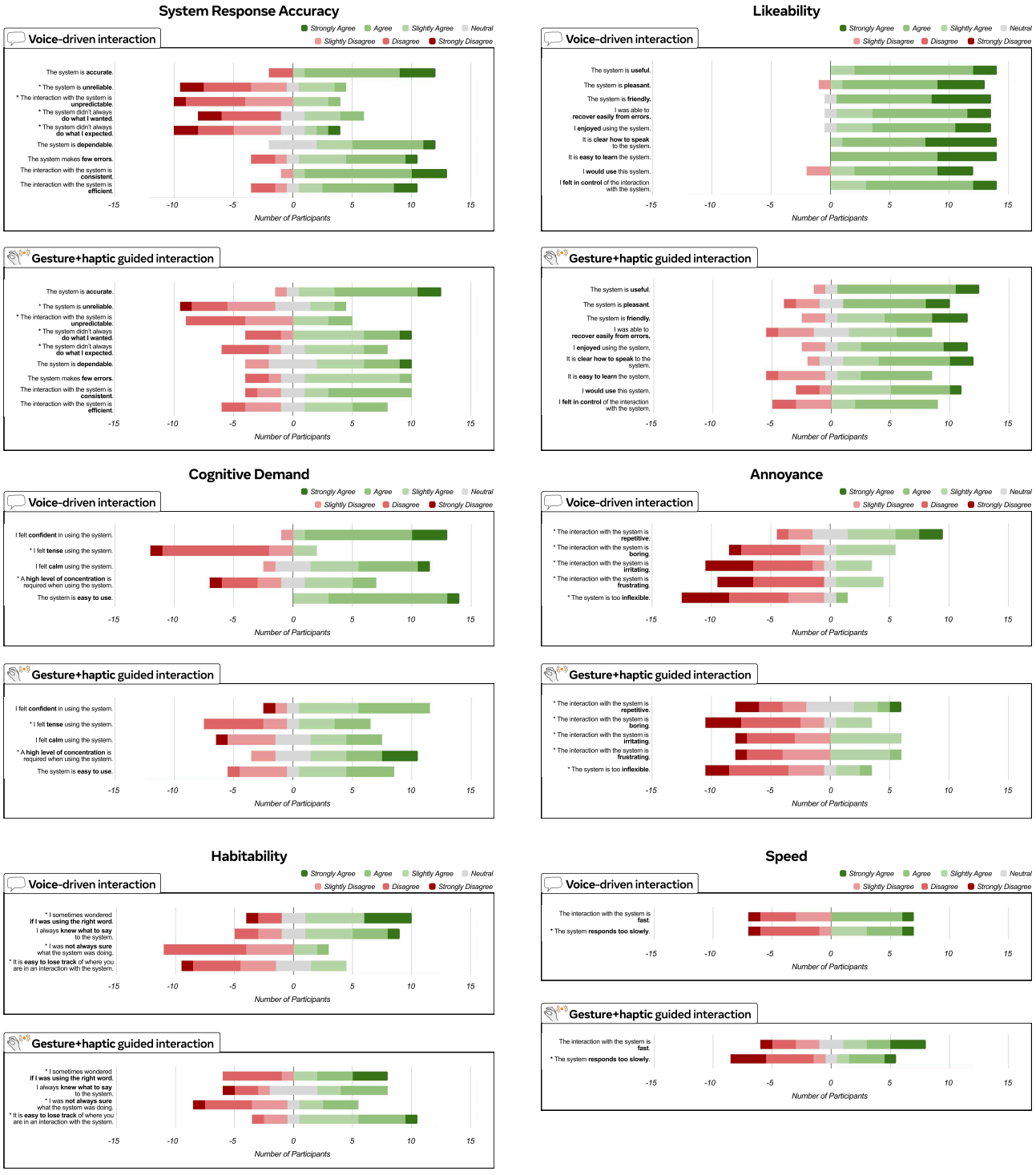

Figure 16: *Subjective Assessment of Speech User Interfaces* Questionnaire Results. We visualize all 14 participants' Likert responses as *diverging barcharts with neutrals split*. Questionnaire items with (*) indicate negatively-phrased questions where a disagreeing score is favorable (e.g., "The system is unreliable").

	Mean	Standard Deviation	Minimum	Maximum
System Response Accuracy (Voice-Driven)	5.087	1.0	3.667	6.778
System Response Accuracy (Gesture+Haptic-Guided)	4.484	0.892	2.667	5.889
Likeability (Voice-Driven)	6.016	0.442	5.333	6.778
Likeability (Gesture+Haptic-Guided)	5.071	1.07	3.111	6.444
Cognitive Demand (Voice-Driven)	5.329	0.795	3.8	7.0
Cognitive Demand (Gesture+Haptic-Guided)	4.114	1.055	2.2	5.4
Annoyance (Voice-Driven)	4.857	1.191	2.6	6.6
Annoyance (Gesture+Haptic-Guided)	4.743	1.273	2.6	7.0
Habitability (Voice-Driven)	4.393	0.848	3.0	5.75
Habitability (Gesture+Haptic-Guided)	3.821	1.443	1.5	6.25
Speed (Voice-Driven)	4.5	2.0	1.0	7.0
Speed (Gesture+Haptic-Guided)	4.179	1.957	1.0	6.5

Table 3: Descriptive Statistics for the Subjective Assessment of Speech User Interfaces (SASSI) Questionnaire. Ratings are on a 7-point Likert scale, where higher scores indicate more favorable responses (1=Strongly Disagree, 4=Neutral, 7=Strongly Agree).

	Test Statistic (Z)	p-value	Effect Size (r)
System Response Accuracy	1.922	0.059	0.514
***Likeability**	2.731	**0.007**	0.73
***Cognitive Demand**	2.605	**0.01**	0.696
Annoyance	0.314	0.779	0.084
Habitability	1.381	0.173	0.369
Speed	0.804	0.43	0.215

Table 4: Statistical Analysis for the Subjective Assessment of Speech User Interfaces (SASSI) Questionnaire. The questionnaire data was analyzed via a Wilcoxon signed-rank test for two paired samples. Due to the small sample size, we also report the effect size to assess the magnitude of observed differences. For all statistically significant differences ($p < 0.05$), we observed a large effect size ($|r| > 0.5$).

Dimensions where the Voice-Driven condition scored significantly higher than the Gesture+Haptic-Guided condition (Likeability and Cognitive Demand) are marked with (*).

A.3 LLM Prompts for Voice Interface Functionality

A.3.1 Informational Audio Assistant. This was the primary agent we used to generate responses to users' prompts. To help GPT-4o maintain the context of the conversation with follow-up requests, we passed in a truncated conversation history (i.e., the most recent user prompt and system response) and optional stylistic instructions (e.g., "`Respond in a scientific manner`").

System Prompt:

```
Your role is an Informational Audio Assistant that responds to users' queries. You will receive a user prompt.
You may also receive a conversation history and stylistic instructions to apply to your response.

#Guidelines
Your responses will be converted to speech and read out to the user via speakers on their smart glasses.
Follow these guidelines to ensure users can easily parse and remember your responses:

*Concise responses (1-3 sentences)
*Predictable sentence structure: write the key idea in the 1st half of the response, add relevant details in
the 2nd half.
*Tailor your response to the depth of conversation history.
>If the conversation history is empty, return a high-level overview of the topic.
>If the conversation has been ongoing, present more detailed responses.

#Return your output in JSON format:
{ "response": string (1-3 sentences) }
```

User Prompt:

```
Suggest 3 follow-up topics for the current topic: <insert question>

Avoid overlapping in the topic areas covered by previous follow-ups:
<insert history>

Return your output in JSON format:
{ "followUps": [string, string, string] }
```

A.3.2 Response Extender. To enable the gesture-driven **`Tell Me More`** functionality that generates more details at the end of the system's current response, we created a `Response Extender` agent that uses the same System Prompt as the Informational Audio Assistant.

User Prompt:

```
Generate 2-3 extra sentences to your previous response that go deeper into the current topic.

#Guidelines
Go down a rabbit hole (get increasingly detailed). Do *NOT* jump to a new topic.

#Response: <insert response>

#Return your output in JSON format:
{ "responseExtension": string }
```

A.3.3 Keyword Identifier. To enable haptic nudges interleaved with system speech to highlight keywords for users to **`Go Deeper`** into, we created a `Keyword Identifier` agent that identifies a specified number of keywords in the system's current response. We instruct GPT-4o to extract approximately 1 keyword for every 10 words in the system response.

System Prompt:

```
Your role is a Keyword Identifier that extracts interesting keywords for a Voice Assistant User to explore. You will
receive a string of the Voice Assistant's current response.

#Guidelines
```

*Keywords should represent new topics that are interesting for users to explore. They should NOT replicate the main topic of the user's query.
*Keywords should NOT be the first or last word of the sentence. Choose keywords that are evenly distributed throughout the sentence.
*Do not repeat keywords.

#Return your output in JSON format:
{ "keywords": [string, string . . .] }

User Prompt

Identify *at most <insert num keywords>* in the Voice Assistant Response. Choose unique keywords that are evenly distributed throughout the sentence and *DO NOT* choose the first word of the sentence.

#Response: <response>

#Return your output in JSON format:
{ "keywords": [string, string . . .] }

A.3.4 Follow-up Generator. To enable the gesture-driven **Play Follow-ups** functionality, we developed a Follow-up Generator agent that returns three subtopics for the user to explore. We pass in the history of previously suggested follow-up topics to avoid repetition.

System Prompt:

Your role is a Follow-up Topic Generator that suggests subtopics for a Voice Assistant User to explore.

#Input:
You will receive the user's current topic and a list of previous follow-up topics you suggested. Suggest 3 topics that derive naturally from the user's original question and strive to help the user explore different dimensions of topic's design space (e.g., discussing a country's culture in terms of language, food, clothing, etc.)

#Guidelines
1. Concise responses (5 words max) phrased as topics rather than questions (e.g., American food)
2. Avoid overlaps in topic areas for your current and previous follow-up questions. If the user chooses to discuss multiple of these questions, they should not hear repeated information.
3. Make sure you can objectively answer or discuss these questions based on your training data. (If the user asks the follow-up question, you must not hallucinate).

#Return your output in JSON format:
{ "followUps": [string, string, string] }

User Prompt

Suggest 3 follow-up topics for the current topic: <insert question>

Avoid overlapping in the topic areas covered by previous follow-ups:
<insert history>

Return your output in JSON format:
{ "followUps": [string, string, string] }