Published: 14 October 2025

CHItaly 2025: CHItaly 2025: 16th Biannual Conference of the Italian SIGCHI Chapter
October 6 - 10, 2025
Salerno, Italy

POSTER

Enabling Voice-based Direct Manipulation in Intent-oriented Interaction Paradigms

LAURA COLAZZO, Politecnico di Milano, Milan, MI, Italy

MARISTELLA MATERA, Politecnico di Milano, Milan, MI, Italy

CHItaly '25: Proceedings of the 16th Biannual Conference of the Italian SIGCHI Chapter (October 2025)
https://doi.org/10.1145/3750069.3755951
ISBN: 9798400721021

Enabling Voice-based Direct Manipulation in Intent-oriented Interaction Paradigms

Laura Colazzo
Department of Electronics, Information and Bioengineering
Politecnico di Milano
Milano, Italy
colazzolaura@gmail.com

Maristella Matera
Department of Electronics, Information and Bioengineering
Politecnico di Milano
Milano, Italy
maristella.matera@polimi.it

Abstract

This paper presents new interaction mechanisms for voice-based direct manipulation that aim to extend the conversational interaction with LLMs and, more generally, with intent-driven user interfaces. It discusses the need for these extensions and illustrates the proposed approach and prototype, and a preliminary user study.

CCS Concepts

• **Human-centered computing** → **Accessibility systems and tools**; **Interaction paradigms**; **Natural language interfaces**.

Keywords

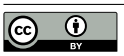
Intent-driven UIs, Voice-based direct manipulation, LLM interfaces, Accessibility

ACM Reference Format:
Laura Colazzo and Maristella Matera. 2025. Enabling Voice-based Direct Manipulation in Intent-oriented Interaction Paradigms. In *CHItaly 2025: 16th Biannual Conference of the Italian SIGCHI Chapter (CHItaly 2025), October 06–10, 2025, Salerno, Italy*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3750069.3755951

1 Introduction

With the emergence of the intent-based paradigm in human-LLM interactions, new usability challenges arise [6]. Ensuring the user intent is accurately captured from a prompt in natural language remains a challenge [5], especially when prompt refinement must be applied iteratively. Usability problems are more pronounced when situational or permanent disabilities require voice interaction [8].

To streamline prompting in strictly conversational LLM GUIs, e.g., ChatGPT, the integration of interaction mechanisms based on *direct manipulation principles* [9] has been proposed [3–5, 7]. Although still in its early stages, the extension of direct manipulation principles to voice-based interactions with LLMs has been explored [2]; however, the paradigm still lacks rigorous validation. As a first step toward proper validation, we introduce a prototype integrating the voice-based direct manipulation paradigm and discuss the result of a preliminary study involving 9 participants.



CHItaly 2025, Salerno, Italy

ACM ISBN 979-8-4007-2102-1/25/10
https://doi.org/10.1145/3750069.3755951

2 Approach

The new mechanisms for voice-based direct manipulation are centered on the definition of an *edit mode*. Conceptually analogous to the canvas area of some visual LLM interfaces [3, 5, 7], it is accessed using a dedicated vocal command (e.g., "*Enter edit mode*"). Being separated from the main chat with the model, the editing space is used to carry out a series of LLM-assisted transformations to a selected prompt or response, while keeping the main chat clean and concise—a requirement emerged from previous studies [8].

In edit mode, user requests to modify prompts can refer to: ***(i) Global editing actions*** (Figure 1): they affect the target message in its entirety (e.g., change its tone). Any time a request of this kind is issued to the LLM, the full modified message is read aloud. In the case of particularly long messages, instead, only a summary of the modifications applied is provided. ***(ii) Localized editing actions*** (Figure 2): they affect only a restricted portion of the message (e.g., replace a word with a synonym). To perform this class of actions in an unambiguous way, the user must be able to both *traverse* the message using a finer-grained and adjustable granularity (e.g., at sentence level) to reach the portion of interest, and to *select* it as a target for transformations. This behaviour can be achieved through navigation patterns [2], implemented within a *navigation mode* that can be activated using a vocal command (e.g., "*Enter navigation mode*"). Users can optionally specify the desired navigation granularity, otherwise the most appropriate one is selected based on the length of the message. In navigation mode, users can also access helper commands that facilitate message traversal. Landing on a given node triggers a series of actions: the node's identifier is read aloud, followed by its content; then, the user receives suggestions for possible actions. At this stage, any editing request considers the current node as the target.

We developed a prototype integrating these mechanisms, which consists of a React front-end embedding Speech-To-Text and Text-To-Speech modules, and a Python back-end, integrating a predefined logic with calls to OpenAI's models. Every user request is processed according to a *state machine*, and is routed to the most relevant *state handler*. Such handlers leverage the LLM to process incoming requests, thus enabling the execution of global and local editing actions on the target message. A detailed description of the implemented voice interaction patterns can be found in [1].

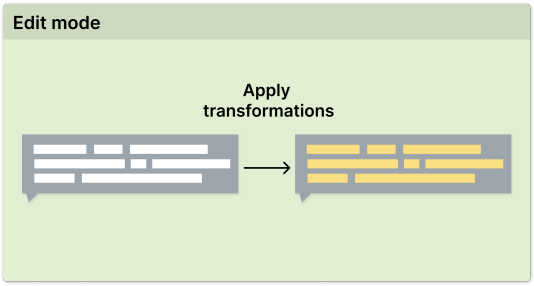


Figure 1: Global editing actions

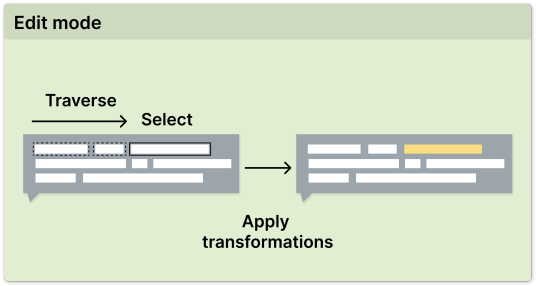


Figure 2: Localized editing actions

3 Preliminary validation

In a preliminary validation[1], 9 participants (6 self-identified as female, 3 as male; ages from 24 to 55 years (M=30, SD=10)) were observed while using the prototype, following a thinking-aloud protocol. None of them reported visual disabilities: while the proposed paradigm was originally motivated by a study involving BVI individuals [8], the evaluation focused on the proposed voice interaction mechanisms independently of sole voice reliance.

Each session lasted approximately 1.5 hours. After completing a questionnaire on demographic data and prior experiences with vocal assistants, participants were given time to familiarize themselves with the prototype and were later observed while completing a set of tasks involving the composition of an email through voice using our prototype. A post-session survey, including questions from SUS, NASA-TLX, TAM, was filled out by each participant to assess the perceived usability, workload, and technology acceptance, respectively. Two open-ended questions also asked about the most liked and disliked aspects. Finally, a semi-structured interview elicited discussion on key aspects of the experience.

3.1 Quantitative Results

The mean SUS score was 86.9 (SD=14.2) on a 0-100 scale, indicating a high level of perceived usability. The mean NASA-TLX score was 35.9 (SD=12.6) on a 0-100 scale, suggesting that participants found the tasks moderately demanding. In particular, *mental demand* was the most critical during task execution, given both the higher scores received (M=4.3 on a 0-10 scale)—indicating a greater perceived cognitive workload—and the greater variability of responses (SD=2.5 on a 0-10 scale)—suggesting that participants experienced mental demand differently. The TAM responses highlighted a high level of perceived acceptance, with notable results in terms of system *learnability* and *ease of use* (4.8 and 4.7 on a 5-point scale, respectively). Their low standard deviations (SD=0.4 and SD=0.5, respectively) also suggest a strong agreement among participants.

3.2 Qualitative Results

The analysis of the users' opinions highlighted that overall the system is perceived as simple, intuitive, and well-designed. Some concerns were raised on the transitions to different modalities (e.g., from the *edit mode* to the *navigation mode*). For example, for the navigation mode, the majority of users assumed its functions were implicit or forgot it existed. Additional suggestions for improvement also emerged (a detailed discussion can be found in [1]):

- **Voice shortcuts** are convenient for recurrent tasks, but naming/recall can be challenging.
- **Numbered options** serve as an anchoring mechanism.
- Preferences for **feedback & summarization** (e.g., full readbacks vs. summaries) are subjective and vary with the task.
- **Voice vs. screen:** voice is preferred for quick actions; screen is preferred for long or critical texts.
- **Delegation & mixed modality:** Delegating tasks is accepted only with user confirmation; voice-plus-visual interaction is valued.

4 Conclusion

This paper has presented preliminary steps in the definition of a voice-based direct manipulation paradigm for human-LLM interaction. Although the conducted study revealed a positive attitude for the new interaction mechanisms, critical reflections also emerged from the participants' qualitative feedback. We therefore plan to organize further investigations to inform next design iterations.

Acknowledgments

This research is supported by the Italian Ministry of University and Research under grant PRIN 2022 "PROTECT" (imPROving ciTizEn inClusivity Through Conversational AI)". CUP: H53D23008150001.

References

[1] Laura Colazzo. 2025. *Voice-based Direct Manipulation in Intent-driven User Interfaces.* Technical Report. Master Thesis, Politecnico di Milano.
[2] Laura Colazzo, Emanuele Pucci, and Maristella Matera. 2025. Voice-based Direct Manipulation to Foster Inclusion in Intent-driven User Interfaces. In *IS-EUD 2025 - Workshops, Work in Progress Demos and Doctoral Consortium (CEUR, Vol. 3978).* https://ceur-ws.org/Vol-3978/short-s2-03.pdf
[3] Google. 2025. New ways to collaborate and get creative with Gemini. https://blog.google/products/gemini/gemini-collaboration-features/
[4] Reuben Luera, Ryan A. Rossi, Alexa Siu, and et al. 2024. Survey of User Interface Design and Interaction Techniques in Generative AI Applications. arXiv:2410.22370 [cs.HC] https://arxiv.org/abs/2410.22370
[5] Damien Masson, Sylvain Malacria, Géry Casiez, and Daniel Vogel. 2024. DirectGPT: A Direct Manipulation Interface to Interact with Large Language Models. In *Proc. of CHI 2024.* ACM, New York, NY, USA, 1–16.
[6] Jakob Nielsen. 2023. AI: First New UI Paradigm in 60 Years. https://www.nngroup.com/articles/ai-paradigm/
[7] OpenAI. 2024. Introducing Canvas. https://openai.com/index/introducing-canvas/
[8] Emanuele Pucci, Ludovica piro, Salvatore Andolina, and Maristella Matera. 2024. From Conversational Web to Inclusive Conversations with LLMs. In *Proc. of AVI '24.* ACM, Article 87.
[9] Shneiderman. 1983. Direct Manipulation: A Step Beyond Programming Languages. *Computer* 16, 8 (1983), 57–69.

[1] Authorized by the ethical board of Politecnico di Milano (authorization no. 58/2024).