Leveraging Pronoun Disambiguation in Multimodal Interaction for Contextual Understanding of Voice Assistant Queries

Septon, Thibaut; Leclercq, Theo; Dumas, Bruno

Published in:
IUI 2025 - Companion Proceedings of the 2025 International Conference on Intelligent User Interfaces

DOI:
10.1145/3708557.3716362

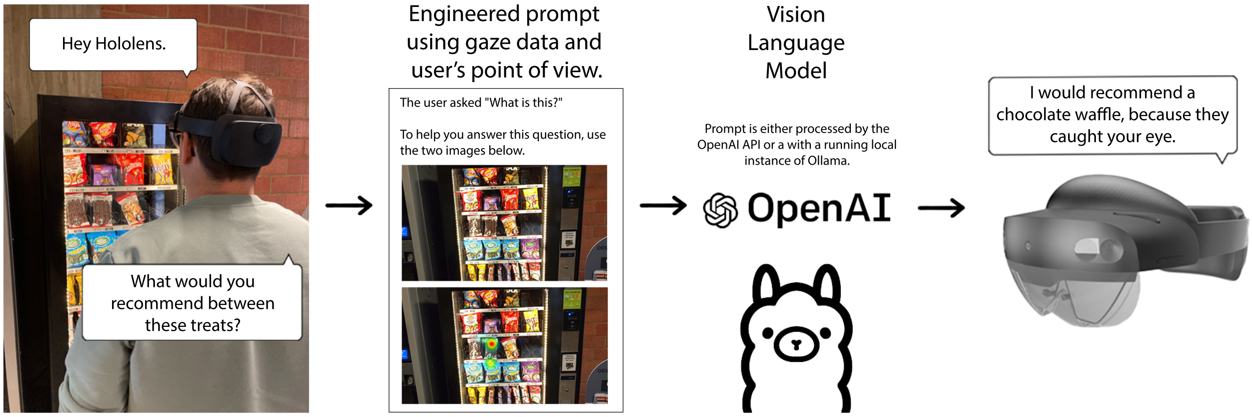
Publication date:
2025

Leveraging Pronoun Disambiguation in Multimodal Interaction for Contextual Understanding of Voice Assistant Queries

Thibaut Septon
thibaut.septon@unamur.be
Université de Namur
Namur Digital Institute
Namur, Belgium

Théo Leclercq
theo.leclercq@unamur.be
Université de Namur
Namur Digital Institute
Namur, Belgium

Bruno Dumas
bruno.dumas@unamur.be
Université de Namur
Namur Digital Institute
Namur, Belgium


Hey Hololens.
What would you recommend between these treats?
Engineered prompt using gaze data and user's point of view.
Vision Language Model
Prompt is either processed by the OpenAI API or a with a running local instance of Ollama.
OpenAI
I would recommend a chocolate waffle, because they caught your eye.

The user asked "What is this?"

To help you answer this question, use the two images below.

Figure 1: Overall workflow and example usage for the VOICE Voice Assistant.

Abstract

Voice Assistants (VAs) are becoming an increasingly important part of our lives. However, most widespread VAs generally fail to take into account the user's spatiotemporal context [11], leading to more descriptive and less natural dialogue. This paper introduces VOICE, an open-source multimodal VA leveraging multimodal interaction and vision-language models to allow for a more flexible and natural communication. Additionally, we present a preliminary user study to evaluate VOICE's ability to understand queries with contextual references.

CCS Concepts

• **Human-centered computing → Natural language interfaces**; *Mixed / augmented reality.*

Keywords

Voice Assistant, Multimodal Interaction, Vision-Language Model, Mixed Reality

ACM Reference Format:
Thibaut Septon, Théo Leclercq, and Bruno Dumas. 2025. Leveraging Pronoun Disambiguation in Multimodal Interaction for Contextual Understanding of Voice Assistant Queries. In *30th International Conference on Intelligent User Interfaces Companion (IUI Companion '25), March 24–27, 2025, Cagliari, Italy.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3708557.3716362

1 Introduction

Voice Assistants (VAs) have become prevalent in our daily lives [17]. With the recent announcements of Meta's Orion, Snapchat's Spectacles, and Android XR, the use of speech interfaces is expected to grow even further. However, forty years after Bolt's seminal work "Put that there" [4], integrating human speech within a User Interface (UI) remains a difficult task. Human communication is inherently multimodal [19], and the English language, in particular, relies heavily on pronouns [5, 7, 12] and the use of complementary modalities (e.g. gaze or gestures) to disambiguate them [6, 8, 15, 20]. As the most commonly deployed VAs struggle with spatiotemporal contextual information [11], the use of deictic pronouns (i.e. referencing contextual elements, such as "this", "there", or "here") presents multiple difficulties. This problem, known as the *pronoun disambiguation* problem [7, 11], affects the naturalness of VA interaction [2].

With recent advancements in the Artificial Intelligence (AI) field, the development of Large Language Models (LLMs) with multimodal capabilities has opened new possibilities for VA interfaces. Specifically, the advent of Vision-Language Models (VLMs) enables

IUI Companion '25, March 24–27, 2025, Cagliari, Italy
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1409-2/25/03
https://doi.org/10.1145/3708557.3716362

machine to process and reason over visual data in conjunction with textual input, making them possibly well-suited to address the challenges posed by spatiotemporal context in human to VA communication.

We introduce VOICE Operating in Contextual Environments (VOICE), an open-source multimodal VA capable of disambiguating multiple pronouns by taking into account the user's context and visual attention. Multimodal interaction has demonstrated flexibility in the allowed user interaction [16], and VLM has exhibited strong reasoning capabilities with images. By combining both, VOICE seeks to explore their potential to enhance VAs. Our preliminary user study shows that VLM additional input capability opens a lot of exciting possibilities for VAs by improving their reasoning capability over environmental contextual information while allowing users to use more flexible queries.

2 Related Work

To resolve pronoun ambiguity, previous works have relied solely on speech through conversations between the user and the VA [13, 21]. Others have used complementary modalities such as touch [9], gaze [14], or pointing gestures [18] to determine user's attention. GazePointAR [10, 11] is a Head-Mounted Display (HMD)-based VA that integrates gaze and pointing gestures. Their VA allows for a more natural conversation as it can comprehend contextually ambiguous queries such as "What is this?" or "Can you recommend something from here?", making interactions more intuitive and context-aware. The latest version of their system, as presented in [11], first listens for the user's query and then waits for the user to stand still while pointing and gazing for the system to take a screenshot through the HMD cameras. The screenshot and the deictic modalities are then subsequently analyzed using an object detection model, and a prompt containing the user's query and a textual description of the user's environment (i.e. what he gazed and pointed at) is then forwarded inside a LLM. The generated response is then read aloud by the HMD.

This work builds on the ideas presented by Lee et al. [10, 11] and explores the use of VLMs for VAs through the implementation of VOICE. The use of a VLM distinguishes our work from GazePointAR [11] and is expected to improve two of its identified problems: 1) an improved reasoning capability since the scene is not described but perceived by the model, and 2) the capture of gaze over time for more natural queries.

3 A Context-Aware Multimodal Voice Assistant

A typical interaction with VOICE is illustrated inside Figure 1. When using VOICE, the user starts by saying "Hey glasses". The application will instantly capture a screenshot of the user's Point of View (POV) through the HMD's camera, listen to their subsequent query, and begin tracking their eye gaze continuously until the query is complete. The application then generates a textual response by forwarding these inputs into a VLM and reads it aloud.

3.1 Architecture

The VOICE software architecture follows a client-server model. The client can be any HMD-based application responsible for capturing a screenshot, the user's query and recording eye gaze during the query's duration. The client then communicates with the server using either an HTTP or a WebSocket connection. All source codes are available under the MIT license for reproducibility purposes[1].

We implemented a client application for the HoloLens 2 using *Unity 2022.3.51f* and Microsoft's *MRTK 3*. The user interacts with VOICE by saying "Hey Hololens", the client then takes a screenshot of user's POV and listens for the user to complete their query. While listening, it logs the user's eye gaze data continuously, over the entire period of the query. Each recorded query is then sent to the VOICE server.

The server is written in Typescript and works with *Node v18.20.4*. Upon receiving a query, the server constructs an empirically determined prompt, following Lee et al. [11]'s recommendation for better results. The prompt is built using user's spoken query, their POV and a modified POV that includes their visual attention (see section 3.1.1). It is then processed using a specified pre-prompt VLM (by default, OpenAI's *gpt-4o*). Generated text output is then sent back to the client. As VLM have shown greater reasoning capability [1, 22], we expect VOICE architecture to outperform GazePointAR's [11] textual approach.

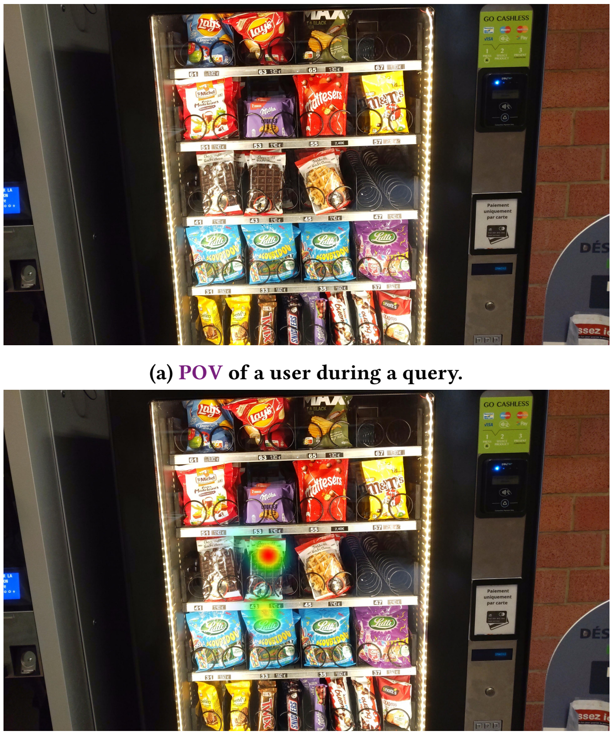
3.1.1 Continuous Gaze Data Representation. In human communication, gaze is directed directly towards referenced objects [8, 15, 20] a key signal used to infer attention and intention, providing insight into what others are focusing on or considering [6]. Therefore, VOICE architecture relies on gaze to resolve pronouns in user queries. As VOICE aims to deliver a natural user experience, with no unnatural forced behavior upon the user, the system records gaze continuously (as suggested by Lee et al. [11]), rather than discretely, during the entirety of any user query. Then, it maps the recorded 3D vectors into their pixel coordinates given the user's POV screenshot taken at the beginning of the query. Subsequently, the user POV is processed by overlaying a heatmap that visually represents where the user focused their gaze during the query (see Figure 2b). This modified POV, enriched with gaze information, serves as a second input to the VLM, allowing it to incorporate the user's visual attention as part of its reasoning process.

4 Preliminary User Study

We gathered 5 participants and asked them each separately to perform 6 pre-defined context-sensitive queries (see Appendix A). Each query was designed to be ambiguous if not supported by an additional deictic modality, with some containing multiple pronouns. Participants were told how to activate VOICE (i.e. the wake up sentence), and were instructed to ask the queries as written, with no further instructions on the interaction. All participants were between 24 and 31 years old ($\mu = 27.2$). All reported having used VAs previously. Through this short evaluation, we aimed to ensure VOICE's ability to disambiguate pronouns and reason over contextual data.

Out of the 30 recorded queries, 22 were correctly disambiguated and understood by VOICE, which proceeded by giving a satisfactory answer. These results are promising and demonstrate the system capability to disambiguate context-aware queries even when containing multiple pronouns. The mean process time per query was

[1] https://github.com/tsepton/voice



(a) POV of a user during a query.

(b) The corresponding user's visual attention.

Figure 2: Example of a recorded POV and its corresponding gaze heatmap.

9.98*s*±1.64. The designed queries left two participants confused at first, as how they should communicate deictic references to the system. However, all independently understood that the system relied on eye gaze[2]. After completing the queries, all participants reported intentionally using their gaze to aid in disambiguation. However, they did not find this use cumbersome, nor did they report that it felt unnatural.

5 Future Works and Limitations

While Section 4 demonstrates that our gaze representation is effective, further adjustments are likely needed. Currently, the heatmap generated from the Hololens eye-tracking data is inaccurate. A better calibration or a more precise eye-tracker could likely benefit the resulting performance. Additionally, our implementation does not account for head movements, which may lead to drift in the data. As suggested by Aziz and Komogortsev [3], incorporating head movement compensation would also likely improve performance.

One participant expressed their will to use pointing gestures and was disappointed to learn later that the system relied solely on gaze as the modality for processing such inputs. When designing VOICE, pointing gestures were excluded as they can be considered an additional effort and/or feel uncomfortable in public settings [11]. Nevertheless, it could enhance the VA reasoning capability but its internal representation should be addressed.

Using a VLM instead of a LLM is expected to enhance the VA reasoning capabilities, but the community could benefit from a comparative study of VOICE with other approaches. Furthermore, the VA will inevitably inherit any limitations present in the chosen model. For example, *gpt-4o* is equipped with safeguards that prevent it from revealing a person's identity. To address such constraints, it is necessary to tackle the limitations inherent in the chosen model. As VOICE is open source, we have made it easy to switch models, including using a self-hosted one through Ollama[3].

While the VLM allows for a deep understanding of the user's environmental context, it currently does not incorporate temporal context, user-specific information (e.g., personal details such as their agenda) nor does it have an internet access. Future research should explore ways to enable the VLM to access past events. However, it is crucial to carefully consider the privacy implications involved.

6 Conclusion

We presented VOICE, a context-aware multimodal VA capable of understanding flexible queries through the capture of the user's gaze. A preliminary user study involving 5 participants assessed a VLM can effectively disambiguate multiple pronouns. Our work shows that VLMs are promising for handling complex, context-sensitive queries, enabling a more nuanced understanding of the user's environment.

Acknowledgments

This version is the preprint version. The authors of this paper would like to express their gratitude to the participants of the user study. Thibaut Septon is supported by the OPTIMIS project by Pôle MecaTech under grant n°8564. Théo Leclercq is supported by the Service Public de Wallonie Recherche under grant n°2310144-WAL4XR.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems* 35 (2022), 23716–23736.

[2] Tawfiq Ammari, Jofish Kaye, Janice Y. Tsai, and Frank Bentley. 2019. Music, Search, and IoT: How People (Really) Use Voice Assistants. *ACM Trans. Comput.-Hum. Interact.* 26, 3, Article 17 (April 2019), 28 pages. https://doi.org/10.1145/3311956

[3] Samantha Aziz and Oleg Komogortsev. 2022. An Assessment of the Eye Tracking Signal Quality Captured in the HoloLens 2. In *2022 Symposium on Eye Tracking Research and Applications* (Seattle, WA, USA) *(ETRA '22)*. Association for Computing Machinery, New York, NY, USA, Article 5, 6 pages. https://doi.org/10.1145/3517031.3529626

[4] Richard A. Bolt. 1980. "Put-that-there": Voice and gesture at the graphics interface. *SIGGRAPH Comput. Graph.* 14, 3, 262–270. https://doi.org/10.1145/965105.807503

[5] Donna Byron and James Allen. 1998. Resolving demonstrative anaphora in the TRAINS93 corpus. (1998).

[6] Herbert H. Clark and Meredyth A. Krych. 2004. Speaking while monitoring addressees for understanding. *Journal of Memory and Language* 50, 1 (2004), 62–81. https://doi.org/10.1016/j.jml.2003.08.004

[7] Albert T. Corbett and Frederick R. Chang. 1983. Pronoun disambiguation: Accessing potential antecedents. *Memory & Cognition* 11, 3 (01 May 1983), 283–294. https://doi.org/10.3758/BF03196975

[8] Ross Flom, Kang Lee, and Darwin Muir. 2017. *Gaze-following: Its development and significance.* Psychology Press.

[2] This was likely influenced by the Hololens requiring calibration for eye gaze when worn for the first time.

[3] https://ollama.com/

[9] Jaewook Lee, Sebastian S. Rodriguez, Raahul Natarrajan, Jacqueline Chen, Harsh Deep, and Alex Kirlik. 2021. What's This? A Voice and Touch Multimodal Approach for Ambiguity Resolution in Voice Assistants. In *Proceedings of the 2021 International Conference on Multimodal Interaction* (Montréal, QC, Canada) *(ICMI '21)*. Association for Computing Machinery, New York, NY, USA, 512–520. https://doi.org/10.1145/3462244.3479902

[10] Jaewook Lee, Jun Wang, Elizabeth Brown, Liam Chu, Sebastian S. Rodriguez, and Jon E. Froehlich. 2023. Towards Designing a Context-Aware Multimodal Voice Assistant for Pronoun Disambiguation: A Demonstration of GazePointAR. In *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23 Adjunct)*. Association for Computing Machinery, New York, NY, USA, Article 92, 3 pages. https://doi.org/10.1145/3586182.3615819

[11] 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 *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 408, 20 pages. https://doi.org/10.1145/3613904.3642230

[12] Geoffrey Leech, Paul Rayson, et al. 2014. *Word frequencies in written and spoken English: Based on the British National Corpus.* Routledge.

[13] Toby Jia-Jun Li. 2020. Multi-Modal Interactive Task Learning from Demonstrations and Natural Language Instructions. In *Adjunct Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '20 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 162–168. https://doi.org/10.1145/3379350.3415803

[14] Sven Mayer, Gierad Laput, and Chris Harrison. 2020. Enhancing Mobile Voice Assistants with WorldGaze. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3313831.3376479

[15] Antje S. Meyer, Astrid M. Sleiderink, and Willem J.M. Levelt. 1998. Viewing and naming objects: eye movements during noun phrase production. *Cognition* 66, 2 (1998), B25–B33. https://doi.org/10.1016/S0010-0277(98)00009-2

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

[17] Reza Rawassizadeh, Taylan Sen, Sunny Jung Kim, Christian Meurisch, Hamidreza Keshavarz, Max Mühlhäuser, and Michael Pazzani. 2019. Manifestation of virtual assistants and robots into daily life: vision and challenges. *CCF Transactions on Pervasive Computing and Interaction* 1, 3 (01 Nov 2019), 163–174. https://doi.org/10.1007/s42486-019-00014-1

[18] Yevhen Romaniak, Anastasiia Smielova, Yevhenii Yakishyn, Valerii Dziubliuk, Mykhailo Zlotnyk, and Oleksandr Viatchaninov. 2020. Nimble: Mobile Interface for a Visual Question Answering Augmented by Gestures. In *Adjunct Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '20 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 129–131. https://doi.org/10.1145/3379350.3416153

[19] Matthew Turk. 2014. Multimodal interaction: A review. *Pattern Recognition Letters* 36 (2014), 189–195. https://doi.org/10.1016/j.patrec.2013.07.003

[20] Femke F van der Meulen, Antje S Meyer, and Willem JM Levelt. 2001. Eye movements during the production of nouns and pronouns. *Memory & Cognition* 29 (2001), 512–521.

[21] Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W. Bruce Croft. 2018. Towards Conversational Search and Recommendation: System Ask, User Respond. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (Torino, Italy) *(CIKM '18)*. Association for Computing Machinery, New York, NY, USA, 177–186. https://doi.org/10.1145/3269206.3271776

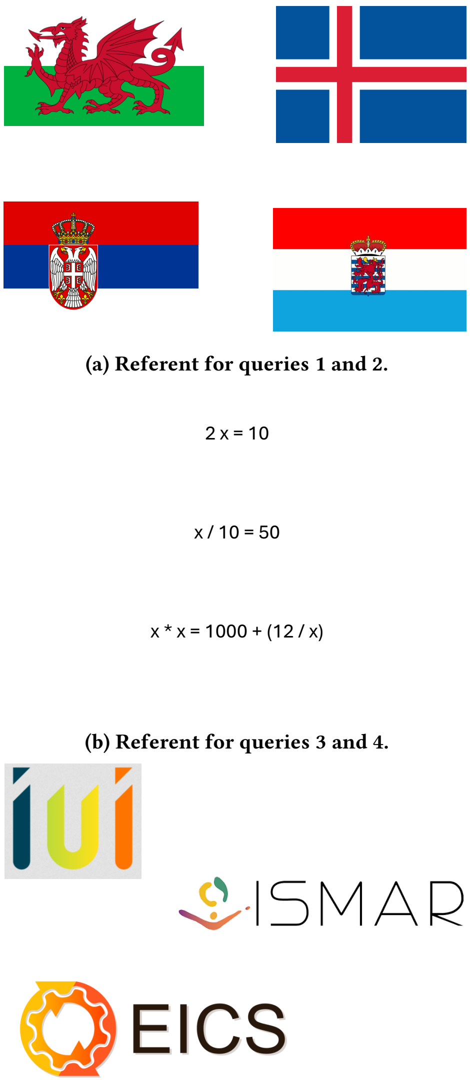
[22] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2024. Multimodal Chain-of-Thought Reasoning in Language Models. arXiv:2302.00923 [cs.CL] https://arxiv.org/abs/2302.00923

A Queries

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009

Original Sentences (French)	English Translation
Quel pays est-ce que c'est?	Which country is this?
Quelle langue est commune à ces deux pays?	Which language is common to these two countries?
Quel est le résultat de ceci [une équation]?	What is the result of this [an equation]?
Le résultat de ceci [une équation] est-il égale à ça [une équation]?	Is the result of this [an equation] equal to that [an equation]?
Connais-tu ce logo?	Do you know this logo?
Sais-tu où ceci prend lieu cette année?	Do you know where this is taking place this year?

Table 1: Ambiguous queries relying on contextual information (see Figure 3) used within the evaluation.



(a) Referent for queries 1 and 2.

(b) Referent for queries 3 and 4.

(c) Referent for queries 5 and 6.

Figure 3: Projected images serving as contextual elements in front of participants during the preliminary user study.