Exploring an Architecture of an Adaptable GenAI-Driven Multimodal User Interface for Third-Party Systems

Tobias Münch
to.muench@muench-its.de
Münch Ges. für IT-Solutions mbH
Lohne (Oldenburg), Germany
Chemnitz University of Technology
Chemnitz, Germany

Martin Gaedke
gaedke@informatik.tu-chemnitz.de
Chemnitz University of Technology
Chemnitz, Germany

Abstract

Nowadays, Generative Artificial Intelligence (GenAI) can outperform humans in creative professions, such as design. As a result, GenAI attracted a lot of attention from researchers and industry. However, GenAI could used to augment humans with a multimodal user interface, as proposed by Ben Shneiderman in his recent work on Human-Centred Artificial Intelligence (HCAI). Most studies of HCAI have mainly focused on greenfield projects. In contrast to existing research, we describe a brownfield software architecture approach with a loosely coupled GenAI-driven multimodal user interface that combines human interaction with third-party systems. A domain-specific language for user interaction connects natural language and signals of the existing system through GenAI. Our proposed architecture enables research and industry to provide user interfaces for existing software systems that allow hands-free interaction.

CCS Concepts

• **Human-centered computing → Interactive systems and tools**; *Interaction techniques*; • **Computer systems organization → Distributed architectures**; • **Computing methodologies →** Artificial intelligence.

Keywords

architecture, interaction, brownfield integration, Domain Specific Language for Interactions, multimodal user interface

1 Introduction

In recent years, industry, the general public, and researchers have become increasingly interested in using Generative Artificial Intelligence (GenAI), such as GPT, to speed up or outsource their tasks. Recently, ChatGPT has released the LLM GPT-4o, which provides a multimodal user interface to interact with its system in a natural human way [9].

This newly implemented Human-Computer Interaction (HCI) approach emphasizes the combination of Intelligence Augmentation (IA) and Artificial Intelligence (AI). Based on Shneiderman and Riedl, this leads to a paradigm shift towards Human-Centred Artificial Intelligence (HCAI) [13, 17]. Traditionally, AI has focused on developing autonomous systems that can replace human labour, such as writing texts or generating images. But it could also help humans improve and enhance their skills by using AI [13, 17]. This paradigm of HCAI emphasises collaborative integration, where technology complements human intellect and skill rather than a substitute [17].

The approach of HCAI supports technologies that synergise with human activities, such as multimodal robot interaction in smart factories proposed by Wang et al. [21]. Central to this evolution is the development of multimodal interfaces, which integrate various forms of communication like speech, touch, eye- or body-movement to create more intuitive and accessible interactions between humans and machines [21]. As we navigate into this new era, the research community needs to ensure that AI development is aligned with human values and designed to meet the needs of users and developers [13, 17].

Our work contributes to HCAI by developing a software architecture concept for GenAI-driven multimodal user interfaces that can extend third-party software systems. We aim to combine different kinds of human interaction with existing software by providing a multimodal adapter (MMA) system that connects both worlds and considers the non-functional requirements (NFR) of users and developers. The transformation of human interaction into system signals is established with a Domain Specific Language (DSL) for user interactions influenced by the work of Li et al. [6].

2 Scenario

Contemporary, we interact daily with several computer systems for business or private use and various end-user devices such as computers, laptops, and smartphones. These systems could be internal or external applications that provide us or the company with value. However, our ability to interact with those specialized systems is limited. Classical input devices or touch gestures provide the primary user interfaces for most business applications. In a specific context, communication with speech or free gestures is nearly unrecognized for industrial applications [12]. These limitations, while significant, also present an opportunity for improvement.

In current development environments, it is rather difficult for software developers to provide a multimodal user interface because of its complex and cost-intensive implementation [1, 3, 24].

Mensch und Computer 2024 – Workshopband, Gesellschaft für Informatik e.V., 01.-04. September 2024, Karlsruhe, Germany

https://doi.org/10.18420/muc2024-mci-ws09-128

Therefore, we focus on software developers' needs and users' expectations. Different NFRs must be met to deliver a successful system in both worlds.

2.1 User Perspective

Based on the work of Oviatt et al., a user could send several natural input signals like eye movement, gestures, and voice towards our proposed interfaces from a wide variety of devices and environments, which have specific requirements and signal errors, such as background noise on an audio-input [10, 11]. The users have various levels of knowledge about computer interaction and operate in several different working environments based on their mental model [4]. Therefore, the user interface has to be straightforward, robust, and safe. In addition, the order sequence of input signals creates a context for a possible system reaction, such as a command to take a picture of this object where my finger points. This context of interactions would enable the system to be used in a human-understandable way.

2.2 Software Developer Perspective

The other perspective is the developer, who will use a Self Development Kit (SDK) to interact with the proposed system. The role of the software developer is to deliver business value in a fast, stable, and secure way into complex production environments [2]. If they cannot figure out how to use the SDK or library clearly in a short period of time, they will move to the next possible solution [18, 20]. The knowledge of the software developer about a third-party library is often limited. Therefore, an understandable interface must be provided for the developer [20].

3 Related Works

Our work builds on previous research on multimodal user interfaces and human interaction through DSL, particularly natural language-to-DSL transformation.

3.1 HCAI and Multimodal User Interfaces

Ben Shneiderman described HCAI as the Second Copernican Revolution, so AI is in the loop around humans, who are the centre of attention [17]. For him, it is a "shift from emulating humans to empowering people" [17].

A multimodal user interface with GenAI is part of Shneiderman's revolution of HCAI because AI is used to help humans interact with machines to enhance themselves [22]. OpenAI has released ChatGPT 4o, ready for multimodal user interaction with its knowledge base. It has several working applications for multimodal interaction with video, gesture, and speech interaction, such as answering external knowledge questions and conversational user interfaces [9]. The idea of a multimodal interface for user agents is familiar and has been mentioned before by Moran et al. [7]. The input signals included natural speech and pen input [7]. ReactGenie proposed a way to interact with react components by natural language and empower the developer to define the usage keywords [6].

The proposed architecture uses GenAI in the context of HCAI to transform human signals into an interaction DSL to interact with different software systems such as ReactGenie for natural language and React [6]. Humans are the main focus of this system, as Shneiderman proposed for HCAI.

In contrast to approaches such as ChatGPT 4o, which use semantic data structures to provide the user with the requested information, we want a loosely coupled GenAI and a loosely coupled DSL to connect various third-party systems to our proposed architecture, enabling developers of multimodal interfaces [9].

3.2 Human Interaction Through a DSL

Firstly, Tablan et al. build a question-answering system to query a knowledge graph by natural language [19]. Yahya et al. mentioned the transformation of natural language into SPARQL for answering questions in the Semantic Web, based on their framework called DEANNA (DEep Answers for maNy Naturally Asked questions) [23]. Based on their work, several prototypes have emerged that use GenAI to transform natural language into a DSL [14, 16]. In addition, Ngonga et al. have shown without GenAI that SPARQL queries can be transformed into natural language [8].

Our approach follows the previously introduced idea of translating natural language into a DSL as input for a third-party interpreter. In our case, we transform natural speech, gestures, or eye movements directly into an interaction DSL or through a signal converter into natural language and then into our DSL. However, our architecture is currently incapable of returning answers in natural language and depends on the reaction of the third-party system.

4 Architecture

This chapter presents an overall architecture to fulfill the described scenarios for a universal and adaptable multimodal system. First, the workflow of user interaction, data, commands, and integration is described from a general perspective. Then, the specific components will be presented in more detail to show their input and output parameters.

4.1 Workflow

The workflow starts with a software developer who wants to integrate an SDK into his primary systems, such as a mobile application or an ERP suite (see Figure 1). To achieve a universal approach, the SDK must be published in a specific format for a specific programming environment, such as an NPM package in JavaScript. For this integration, the SDK provides an adapter that provides an interface for registering voice commands in a natural language analogue to ReactGenie, an approach by Li et al. or Yang et al. [6, 24]. Depending on the programming environment, this registration attaches a callback method or event listener.

After the registration, the data is passed to the trainer, which integrates the new instructions into an essential interaction DSL for training the GenAI (see Figure 1). The trainer also defines standard commands. These do not need to be implemented by the developer. They can be overridden according to the convention over configuration pattern [5]. The training process could be carried out with the help of prompt engineering so that the GenAI could transform natural language into our DSL. For this transformation, the registered commands in natural language are used. In addition, if prompt engineering is used, new commands could be suggested to the generative AI at runtime.

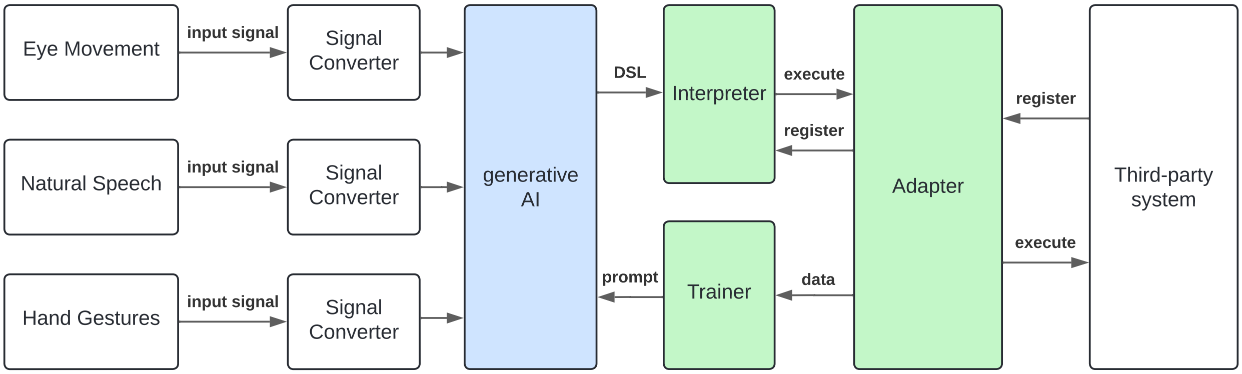

Figure 1: Proposed multi-level architecture for a GenAI-driven multimodal user interface supported by generative artificial intelligence

After the startup and registration phase, the MMA can process user interaction through eye movement, natural speech, and hand gestures (see Fig. 1). A signal converter must convert each interaction signal into a readable format for the GenAI, such as a speech-to-text provider for natural language recognition. Based on the current development of GPT-4o, processing of video streams or other interactive data is also possible [9]. The input parameters for GenAI are transformed to the previously defined DSL.

The interpreter then processes the DSL and links it to the methods previously registered by the adapter. The adapter then executes the commands in the integrated system. Depending on whether the third-party system has been integrated at the user interface or business logic level, the user receives visual feedback.

4.2 Components

As briefly described in the workflow, the proposed MMA system contains several components with a dedicated objective. The signal converter is a component that takes user input signals from a source and performs the vital task of converting them into a format that the GenAI can interpret. For example, if the GenAI is ChatGPT 4, the converter must generate natural text. Other signals, such as a muscle sensor, could extend these input signals. The only requirement is that the signal can be converted into a GenAI-readable format. The GenAI component's objective is to transform an input into the defined DSL, which the Trainer component can enrich on demand. The interpreter uses the DSL to interact with the adapter, which is the glue between the external system and the multimodal interaction.

The GenAI component is primarily designed to facilitate the translation process, converting natural language into the interaction DSL, thereby enabling effective communication with the MMA system. Any service that provides the required privacy policies, an open API such as RESTful web services, adequate performance, and accurate translation of natural language into our to-be-defined DSL could be used. The GenAI provider must provide a natural language interface for user input signals and our trainer component as a minimum requirement.

The core of our application is a to-be-defined interaction DSL, which provides the main interaction patterns for desktop, mobile, and augmented or virtual reality applications. It should be extendable by developers with custom commands or interaction patterns. It connects third-party applications with the interpreter to the natural language of the users. The interpreter component is the controller of the connected third-party system. It translates the incoming DSL and executes registered methods through the adapter component.

5 Challenges

Our proposed architecture is an early design draft for a universal, loosely coupled approach for multimodal user interaction with various challenges that must be addressed. These challenges are categorized into 1) user interaction through GenAI, 2) adequately designed DSL for various user interactions on several platforms, and 3) performance and error rate.

The input signals must be interpreted by a GenAI or converted into a readable format. How can the error rate and mean response time be minimised? Recent research has shown that these are critical metrics for user acceptance and should be managed or prevented [15]. In addition, how to respond to parallel inputs via the multimodal interface and transform them together for the GenAI to provide a context-specific response to the user.

The DSL briefly described above had to be defined so that different interactions could be linked to interaction patterns in a software application such as a web browser. This DSL must be understandable to the GenAI and interpretable for integration. Furthermore, how can the performance be maximised and resource consumption be minimised?

Additionally, an error channel has to be included in our approach, or the third-party system has to be forced to provide an error output because, with sufficient probability, there will be several translation or transformation errors. Finally, how could audio-visual feedback be integrated into the proposed architecture or SDK using device-specific capabilities?

6 Conclusion

In line with Shneiderman's vision of HCAI, we have developed a draft for a GenAI-driven multimodal software architecture. This architecture is a two-way conduit, connecting natural human interactions with existing third-party systems. This connection could be achieved by our proposed DSL for interaction. Our DSL connects human interaction with system signals of third-party systems through GenAI and an Interpreter, which is part of our SDK. This SDK focuses on software developers who want to connect their systems quickly. If we can handle the described challenges, our architecture could be published across multiple platforms, impacting various software products worldwide.

References

[1] Sean Andrist, Dan Bohus, Ashley Feniello, and Nick Saw. 2022. Developing mixed reality applications with platform for situated intelligence. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 48–50.

[2] Joseph D Blackburn, Gary D Scudder, and Luk N Van Wassenhove. 1996. Improving speed and productivity of software development: a global survey of software developers. *IEEE transactions on software engineering* 22, 12 (1996), 875–885.

[3] Dan Bohus, Sean Andrist, Ashley Feniello, Nick Saw, Mihai Jalobeanu, Patrick Sweeney, Anne Loomis Thompson, and Eric Horvitz. 2021. Platform for situated intelligence. *arXiv preprint arXiv:2103.15975* (2021).

[4] John M Carroll and Judith Reitman Olson. 1988. Mental models in human-computer interaction. *Handbook of human-computer interaction* (1988), 45–65.

[5] Nicholas Chen. 2006. Convention over configuration. *h ttp://softwareengineering. vazexqi. com/files/pattern. htm l* (2006).

[6] Karina Li, Daniel Wan Rosli, Shuning Zhang, Yuhan Zhang, Monica S Lam, James A Landay, et al. 2023. ReactGenie: An Object-Oriented State Abstraction for Complex Multimodal Interactions Using Large Language Models. *arXiv preprint arXiv:2306.09649* (2023).

[7] Douglas B Moran, Adam J Cheyer, Luc E Julia, David L Martin, and Sangkyu Park. 1997. Multimodal user interfaces in the Open Agent Architecture. In *Proceedings of the 2nd international conference on Intelligent user interfaces*. 61–68.

[8] Axel-Cyrille Ngonga Ngomo, Lorenz Bühmann, Christina Unger, Jens Lehmann, and Daniel Gerber. 2013. Sorry, i don't speak SPARQL: translating SPARQL queries into natural language. In *Proceedings of the 22nd international conference on World Wide Web*. 977–988.

[9] OpenAI. 2024. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/. (Accessed on 05/15/2024).

[10] Sharon Oviatt. 2007. Multimodal interfaces. *The human-computer interaction handbook* (2007), 439–458.

[11] Sharon Oviatt and Philip Cohen. 2000. Perceptual user interfaces: multimodal interfaces that process what comes naturally. *Commun. ACM* 43, 3 (2000), 45–53.

[12] Goran Paun. [n. d.]. Voice, Gesture And Emotion: The New Frontier In Accessible User Interfaces. https://www.forbes.com/sites/forbesagencycouncil/2023/05/10/voice-gesture-and-emotion-the-new-frontier-in-accessible-user-interfaces/?sh=612cf517578c. (Accessed on 05/26/2024).

[13] Mark O Riedl. 2019. Human-centered artificial intelligence and machine learning. *Human behavior and emerging technologies* 1, 1 (2019), 33–36.

[14] Md Rashad Al Hasan Rony, Uttam Kumar, Roman Teucher, Liubov Kovriguina, and Jens Lehmann. 2022. Sgpt: A generative approach for sparql query generation from natural language questions. *IEEE Access* 10 (2022), 70712–70723.

[15] Dirk Schnelle-Walka. 2010. A pattern language for error management in voice user interfaces. In *Proceedings of the 15th European Conference on Pattern Languages of Programs*. 1–23.

[16] Numair Shaikh, Tavishee Chauhan, Jayesh Patil, and Sheetal Sonawane. 2024. Explicable knowledge graph (X-KG): generating knowledge graphs for explainable artificial intelligence and querying them by translating natural language queries to SPARQL. *International Journal of Information Technology* (2024), 1–11.

[17] Ben Shneiderman. 2020. Human-centered artificial intelligence: Three fresh ideas. *AIS Transactions on Human-Computer Interaction* 12, 3 (2020), 109–124.

[18] Diomidis Spinellis. 2019. How to select open source components. *Computer* 52, 12 (2019), 103–106.

[19] Valentin Tablan, Danica Damljanovic, and Kalina Bontcheva. 2008. A natural language query interface to structured information. In *The Semantic Web: Research and Applications: 5th European Semantic Web Conference, ESWC 2008, Tenerife, Canary Islands, Spain, June 1-5, 2008 Proceedings 5*. Springer, 361–375.

[20] Minaoar Hossain Tanzil, Gias Uddin, and Ann Barcomb. 2024. " How do people decide?": A Model for Software Library Selection. *arXiv preprint arXiv:2403.16245* (2024).

[21] Tian Wang, Pai Zheng, Shufei Li, and Lihui Wang. 2024. Multimodal Human–Robot Interaction for Human-Centric Smart Manufacturing: A Survey. *Advanced Intelligent Systems* 6, 3 (2024), 2300359.

[22] Wei Xu, Marvin J Dainoff, Liezhong Ge, and Zaifeng Gao. 2023. Transitioning to human interaction with AI systems: New challenges and opportunities for HCI professionals to enable human-centered AI. *International Journal of Human–Computer Interaction* 39, 3 (2023), 494–518.

[23] Mohamed Yahya, Klaus Berberich, Shady Elbassuoni, Maya Ramanath, Volker Tresp, and Gerhard Weikum. 2012. Natural language questions for the web of data. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. 379–390.

[24] Jackie Yang, Yingtian Shi, Yuhan Zhang, Karina Li, Daniel Wan Rosli, Anisha Jain, Shuning Zhang, Tianshi Li, James A Landay, and Monica S Lam. 2024. ReactGenie: A Development Framework for Complex Multimodal Interactions Using Large Language Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–23.