Talking to Data Visualizations: Opportunities and Challenges

Gabriela Molina León*
University of Bremen, Germany

Petra Isenberg†
Université Paris-Saclay, CNRS, Inria, LISN, France

Andreas Breiter‡
University of Bremen, Germany

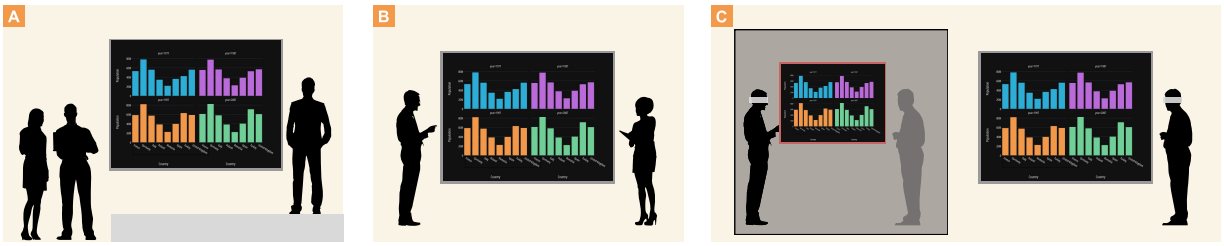

Figure 1: Examples of collaborative scenarios: (a) A presenter giving a lecture to an audience. (b) Two collaborators interacting with data visualizations. (c) Two persons interact through virtual reality. Silhouettes by Gineton Rodrigues [18],

Abstract

Speech is one of the interaction modalities that we increasingly come across in natural user interfaces. However, its use in collaborative scenarios has not yet been thoroughly investigated. In this reflection statement, we discuss the opportunities and challenges of integrating speech interaction in multimodal solutions for collaborative work with data visualizations. We discuss related findings from other research communities and how we could build upon their work to explore and make use of speech interaction for data visualizations in co-located, hybrid, and remote settings.

Index Terms: Human-centered computing—Interaction design; Human-centered computing—Collaborative and social computing systems and tools; Human-centered computing—Visualization—Visualization design and evaluation methods

1 Introduction

Since the launch of ChatGPT in November 2022 [15], the interest of the general public in artificial intelligence (AI) has vastly increased [16]. Among the most notorious features of this chatbot is its ability to process questions in more than 50 languages and to generate answers accordingly. As a result, more than 100 million people have interacted with it worldwide so far [5]. In the context of such technological advancements, we want to reflect on the opportunities and challenges that leveraging natural language interaction may pose for collaborative data visualization. More specifically, we examine the case where natural language is used via voice input and not only via written text. Oral communication plays a critical role in collaborative activities, and thus, speech input can leverage people's existing language skills and provide new ways of interacting with data. In this reflection statement, we highlight the opportunities and challenges that we consider most important for introducing speech interaction into collaborative visual systems.

Most previous work in speech interaction with data visualizations has focused on single-user scenarios [10, 23]. We build on this work by discussing how we could transfer findings on speech interaction for single-user activities for designing collaborative systems. Accordingly, we propose a research agenda to understand better how speech interaction could support or influence collaboration. The opportunities and challenges we discuss are by no means exhaustive. Our goal is to start a discussion on the potential role of speech interaction in different collaboration settings.

*e-mail: molina@uni-bremen.de
†e-mail: petra.isenberg@inria.fr
‡e-mail: abreiter@uni-bremen.de

2 Speech Interaction

Speech interaction refers to the use of voice commands to produce a change or response in a computing system. In the field of human-computer interaction, using natural language to interact is recommended due to its expressiveness and the possibility of supporting interaction by broad audiences [21]. Interacting through written text already works successfully in commercial visualization tools such as Tableau [25]. Moreover, speech is considered an interaction modality that can help make data visualizations more accessible to, for example, blind and low vision users [9]. However, speech commands can be hard to discover [4], so using them efficiently requires a learning phase. Baughan et al. [3] analyzed the failures of voice assistants and found that incorrect command execution, missed triggers, and overcapturing affect user trust.

Prior work on visualization systems for single-user scenarios suggests that the limitations of speech interaction can be addressed by a multimodal interaction design. Kim et al. [10] designed a mobile application that leverages touch and speech interaction to explore personal health data. In their study, participants found speech fast and flexible to make comparison queries and considered that the combination of touch and speech was helpful to refine previous commands. Aurisano et al. [1] proposed to combine speech and mid-air gestures to create and manipulate views on a wall display and found that their study participants combined both modalities efficiently to create multiple views. Srinivasan et al. [23] proposed combining touch, pen, and speech interaction to work with unit visualizations. They found that using direct manipulation and speech interaction facilitated a fluid experience and that participants solved speech recognition errors using touch and pen.

For multi-user scenarios, Tse et al. [26] started exploring the multimodal design space for co-located collaboration combining speech and touch in a tabletop application. The authors concluded that if most interaction techniques are associated with speaking, that may influence how often collaborators talk to each other. Additionally, the authors suggested using speech commands for global mode switches and touch gestures for individual mode switches. However, speech interaction may have different effects and uses in hybrid and remote settings, where people are not necessarily next to each other, may move around vertical displays, and work asynchronously.

3 Opportunities

In the following, we describe the main opportunities that leveraging speech interaction can provide in co-located, hybrid, and remote collaborative work with data visualizations.

3.1 Interacting from any distance

Speech input enables distant interaction, which is relevant in collaborative scenarios involving large vertical displays. People tend to navigate physically by walking in front of such screens, and physical navigation may correlate with user performance [2]. In a study that we conducted recently, participants favored speech over mid-air gestures to interact with data visualizations from afar [12]. While speech can facilitate distant interaction, other modalities can enable interaction up close in a multimodal system. For example, Srinivasan et al. [23] proposed to explore network data on large vertical displays combining direct manipulation (touch and pen) with speech interaction. Such a combination of modalities can be of use not only in co-located scenarios but also in hybrid and remote setups. In those cases, participants may wish to move even while alone in the room (e.g., in a virtual reality scenario as shown in Fig. 1c).

3.2 Requiring only a microphone

Regarding hardware, speech interaction only requires a working microphone to listen to the speech commands. Therefore, users are free to use their hands for interacting with the system in other ways. Speech neither requires looking at a screen nor an additional device, so gaze distractions are not a problem. Regardless of whether the collaboration is happening in a co-located, hybrid, or remote setting, the physical characteristics of the interaction do not change. Additionally, using a microphone is a standard in collaborative commercial platforms, especially when the collaborators need to communicate synchronously. Thus, no extra equipment is necessary. Giving each person access to a microphone in a co-located scenario, however, is recommended to ensure accuracy and to identify each speaker. From a software perspective, speech recognition only demands using dedicated libraries or tools, such as web APIs (e.g., Mozilla's API [14]) or AI-based tools to support a custom vocabulary (e.g., Picovoice [7]). Nowadays, being an AI expert is not a requirement to leverage AI-based approaches. While free and open-source toolkits are still rare, there are already a few options, such as Vosk [6].

3.3 Adapting to speech-heavy activities

Collaborative scenarios can involve not only sensemaking but also other activities, such as presenting and teaching about visualization techniques and tools (e.g., Fig. 1a), where one person is responsible for most of the oral communication. In such cases, we could incorporate speech commands directly into the presentation or teaching content. The presenter or the audience can then use other modalities to interact while the person is talking. Interactive speeches could work in co-located, hybrid, or remote settings, as long as the setup includes a microphone. Although voice assistants are usually activated with a wake word (e.g., "Alexa", "Hey Siri", etc.), using such a phrase is not required. Instead, the system could listen to the conversation, waiting for an opportunity to participate [27], for example, to provide a data fact relevant to the discussion. Studying storytelling techniques may help to propose ways of incorporating speech commands into a presentation or lecture. For example, Shin et al. [22] proposed a related solution by generating natural language narratives to present a sequence of data changes. Their system created textual narratives and animations to highlight temporal changes in a scatterplot. Thus, the presenter could trigger the animations with commands that form parts of their speech and use other interaction modalities for further effects.

3.4 Supporting multilingual interaction

Nowadays, speech interaction tools support the recognition of dozens of languages, which could support collaborative work between people who speak different languages. Previous work shows that people appreciate speech interaction because they can refer orally to concepts that are tedious to specify as data queries through direct manipulation [21]. The option to interact orally in their native language may improve the user experience of the collaborators. Moreover, a system could provide similar flexibility for using expert and non-expert terms to interact orally. For example, a person could say "draw a line that shows if my points are increasing or decreasing" while someone else could say "add a linear trend line".

4 Challenges

We now present the main challenges that researchers and practitioners may face when leveraging speech interaction in the design of collaborative systems.

4.1 Recognition errors

In the context of interactive data visualization, Srinivasan and Stasko [24] reported that study participants sometimes became frustrated while interacting due to speech detection errors. When voice assistants fail to understand speech commands time after time, people tend to trust them less [3]. Providing multimodal interaction may be a solution to this challenge. For example, in the study of Saktheeswaran et al. [19], participants appreciated being able to correct speech recognition and ambiguity errors via touch. Moreover, we are currently witnessing significant improvements in voice recognition systems. The recent launch of deep-learning-based tools such as Whisper [17], Vosk [6], and Picovoice [7] suggests promising advances in speech interaction accuracy.

4.2 Collaboration conflicts

Given that dialogue is a fundamental aspect of collaboration, using speech to interact with the system, and not only with other humans, may disturb the conversation flow. Collaboration partners may hesitate to use speech commands to avoid interrupting each other. However, in our recent elicitation study [12], participants sometimes wished for the system to listen to them and directly map part of their conversation to interactions with the data visualizations. Furthermore, problems may arise during turn-taking, as overlapping speech commands may make speech recognition difficult. As the number of collaborators increases, avoiding overlapping voices becomes harder. In this situation, it is also crucial that the system gives appropriate feedback to help the collaborators understand whether the speech commands are understood and processed.

4.3 Privacy concerns

People may not necessarily be comfortable with a device constantly listening to them, waiting for voice input. Although the wish for privacy-friendly designs can depend on the country and social norms [20], privacy is a primary factor for user acceptance. Moreover, people are protective of saving and making voice recordings of themselves and others available online with cloud services they do not trust. Thus, it is necessary to find alternatives to constant listening and storing voice recordings [11]. Additionally, speech interaction requires talking out loud, which does not allow users to make their interactions private. However, that may not be a problem in asynchronous collaborations. Also, multimodal systems could support private interactions through other modalities to complement speech-based public interactions [26]. For limiting data sharing, there are speech tools that do not require processing voice input online (e.g., Picovoice [7]). They process it on the local machine, which may be a suitable alternative.

4.4 Support for language learners

Although chatbots and voice assistants support multiple languages, we must consider that many people interact with computing systems in English or another language that may not be their first language. Having an accent or not being familiar with the pronunciation of specific words can be an obstacle to speech recognition [13, 24]. Similarly, there are thousands of spoken languages worldwide, and thus, speech interaction is limited by the languages supported by the corresponding software. Machine translation advances may provide more support in the future, but it has limits as well.

5 Research Agenda

Based on the opportunities and challenges mentioned above, we list the research directions we consider most critical and promising for developing scientific knowledge on the role of speech interaction in multimodal and collaborative scenarios.

5.1 Understanding speech in collaboration

As research on conversational user interfaces and multimodal interaction has focused on single-user scenarios, we need to conduct more studies to investigate different collaborative activities to understand how speech interaction can support and affect collaborative work with data visualizations. Similar to how Kim et al. [9] created and published a corpus of questions that participants posed in their study for future speech-based design, collaboration studies could help generate such corpora of questions and dialogues that may be relevant for designing collaborative systems. Collecting these data can help us understand better what tasks are most suitable to perform via speech during collaborative work and what commands are relevant for different collaboration settings and styles.

5.2 Leveraging more design methods

Although current commercial products supporting natural language have known limitations in speech recognition, that should not be an impediment to investigating the potential of speech interaction. Given the latest advancements in AI, natural language technologies may evolve rapidly. We should think beyond the current technological capabilities, as diverse research communities have successfully done. Voice assistant researchers often investigate and propose conversational scenarios through storyboards and Wizard of Oz studies (e.g., [8,27]) to investigate the relevant factors in conversational interface design. We should leverage such methods to inform the design of future visual systems by exploring what could help collaborators most and how.

5.3 Exploring collaborative and multimodal solutions

As mentioned above through different examples, many limitations of speech interaction can be addressed by combining speech with other interaction modalities. Prior work suggests that combining natural language with direct manipulation is a promising solution [23], as well as combining speech commands with mid-air gestures [1]. We should examine how collaboration partners use multiple interaction modalities and how those may influence their strategies and interaction choices. Combining modalities that support up close and distant interaction could help support collaboration at different distances from the visualizations [12].

5.4 Providing voice-based feedback

Natural language interfaces can include not only voice input but also voice output. Given the advancements in voice assistants, we should explore the scenarios in which a system can generate prompts in natural language and be part of the conversation. Zargham et al. [27] already started exploring these scenarios by investigating in what situations a proactive voice assistant could intervene in a decision-making conversation or debate. In the context of presentations and teaching activities, we should consider how a voice assistant could support explanatory visualizations or how it could explain how to read a visualization. Hence, it is worth investigating how the human-assistant interaction could work and how the participation of voice assistants may influence the collaborators. This would be an interdisciplinary endeavour to better understand flaws in communication.

6 Conclusion

This reflection statement discussed the use of speech interaction in collaborative scenarios. We reflected on the opportunities and challenges of incorporating this interaction modality into multi-user interactive systems. Accordingly, we discussed how to leverage speech in diverse collaborative activities and setups involving data visualizations. Given the recent advancements in AI-based approaches for natural language interaction, we should investigate how natural language could facilitate working with data visualizations from any distance, in a group, and even in multiple languages. We should explore this topic from the different perspectives of the related research communities, such as those working on human-computer interaction, natural language processing, and computer-supported cooperative work. Such interdisciplinary efforts will help us better understand whether and how speech could become a valuable interaction modality in collaborative visual systems.

Acknowledgments

This work was partly funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 374666841 – SFB 1342.

References

[1] J. Aurisano, A. Alsaiari, A. Kumar, B. Di, A. J. Eugenio, J. Leigh, and M. Zellner. Ditto: Exploring data in large display environments through speech+mid-air gesture interactions. In *IEEE VIS Posters*, 2022.

[2] R. Ball, C. North, and D. A. Bowman. Move to improve: Promoting physical navigation to increase user performance with large displays. In *Proc. CHI*, pp. 191–200. ACM, New York, 2007. doi: 10.1145/1240624.1240656

[3] A. Baughan, X. Wang, A. Liu, A. Mercurio, J. Chen, and X. Ma. A mixed-methods approach to understanding user trust after voice assistant failures. In *Proc. CHI*. ACM, New York, 2023. doi: 10.1145/3544548.3581152

[4] J. D. Hincapié-Ramos, X. Guo, P. Moghadasian, and P. Irani. Consumed endurance: A metric to quantify arm fatigue of mid-air interactions. In *Proc. CHI*, pp. 1063–1072. ACM, New York, 2014. doi: 10.1145/2556288.2557130

[5] K. Hu. ChatGPT sets record for fastest-growing user base, Feb. 2023.

[6] A. C. Inc. VOSK offline speech recognition API, July 2023.

[7] P. Inc. Train, develop and deploy custom voice features - Picovoice, July 2023.

[8] S. Kernan Freire, E. Niforatos, Z. Rusak, D. Aschenbrenner, and A. Bozzon. A conversational user interface for instructional maintenance reports. In *Proc. CUI*. ACM, New York, 2022. doi: 10.1145/3543829.3544516

[9] J. Kim, A. Srinivasan, N. W. Kim, and Y.-S. Kim. Exploring chart question answering for blind and low vision users. In *Proc. CHI*. ACM, New York, 2023. doi: 10.1145/3544548.3581532

[10] Y.-H. Kim, B. Lee, A. Srinivasan, and E. K. Choe. Data@hand: Fostering visual exploration of personal data on smartphones leveraging speech and touch interaction. In *Proc. CHI*. ACM, New York, 2021. doi: 10.1145/3411764.3445421

[11] N. Malkin, J. Deatrick, A. Tong, P. Wijesekera, S. Egelman, and D. Wagner. Privacy attitudes of smart speaker users. *Proc. on Privacy Enhancing Technologies*, 2019(4), 2019. doi: 10.2478/popets-2019-0068

[12] G. Molina León, P. Isenberg, and A. Breiter. Eliciting multimodal and collaborative interactions for data exploration on large vertical displays, 2023. In submission.

[13] G. Molina León, M. Lischka, W. Luo, and A. Breiter. Mobile and multimodal? A comparative evaluation of interactive workplaces for

visual data exploration. *Computer Graphics Forum*, 41(3):417–428, 2022. doi: 10.1111/cgf.14551

[14] Mozilla and individual contributors. Web speech API, 2023.

[15] OpenAI. Introducing ChatGPT, 2022.

[16] M. Petrova. ChatGPT has made AI the hot new thing in Silicon Valley, and investors are suddenly very interested, 2023.

[17] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, eds., *Proc. ICML*, vol. 202, pp. 28492–28518. PMLR, July 2023.

[18] G. Rodrigues. People silhouettes, 2021.

[19] A. Saktheeswaran, A. Srinivasan, and J. Stasko. Touch? speech? or touch and speech? Investigating multimodal interaction for visual network exploration and analysis. *IEEE Trans. Visual Comput. Graphics*, 26(6):2168–2179, 2020. doi: 10.1109/TVCG.2020.2970512

[20] A. Seiderer, H. Ritschel, and E. André. Development of a privacy-by-design speech assistant providing nutrient information for german seniors. In *Proc. of the EAI International Conference on Smart Objects and Technologies for Social Good*, pp. 114–119. ACM, New York, 2020. doi: 10.1145/3411170.3411227

[21] V. Setlur, S. E. Battersby, M. Tory, R. Gossweiler, and A. X. Chang. Eviza: A natural language interface for visual analysis. In *Proc. UIST*, pp. 365–377. ACM, New York, 2016. doi: 10.1145/2984511.2984588

[22] M. Shin, J. Kim, Y. Han, L. Xie, M. Whitelaw, B. C. Kwon, S. Ko, and N. Elmqvist. Roslingifier: Semi-automated storytelling for animated scatterplots. *IEEE Trans. Visual Comput. Graphics*, 29(6):2980–2995, 2023. doi: 10.1109/TVCG.2022.3146329

[23] A. Srinivasan, B. Lee, and J. Stasko. Interweaving multimodal interaction with flexible unit visualizations for data exploration. *IEEE Trans. Visual Comput. Graphics*, 27(8):3519–3533, 2021. doi: 10.1109/TVCG.2020.2978050

[24] A. Srinivasan and J. Stasko. Orko: Facilitating multimodal interaction for visual exploration and analysis of networks. *IEEE Trans. Visual Comput. Graphics*, 24(1):511–521, 2018. doi: 10.1109/TVCG.2017.2745219

[25] M. Tory and V. Setlur. Do what i mean, not what i say! Design considerations for supporting intent and context in analytical conversation. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 93–103, 2019. doi: 10.1109/VAST47406.2019.8986918

[26] E. Tse, S. Greenberg, C. Shen, C. Forlines, and R. Kodama. Exploring true multi-user multimodal interaction over a digital table. In *Proc. DIS*, pp. 109–118. ACM, New York, 2008. doi: 10.1145/1394445.1394457

[27] N. Zargham, L. Reicherts, M. Bonfert, S. T. Voelkel, J. Schoening, R. Malaka, and Y. Rogers. Understanding circumstances for desirable proactive behaviour of voice assistants: The proactivity dilemma. In *Proc. CUI*. ACM, New York, 2022. doi: 10.1145/3543829.3543834