# Challenges in Making Situated Interactions Accessible to Motor-Impaired Users

**Sai R. Gouravajhala**
Computer Science & Eng.
University of Michigan
sairohit@umich.edu

**Walter S. Lasecki**
Computer Science & Eng.
School of Information
University of Michigan
wlasecki@umich.edu

**Harmanpreet Kaur**
Computer Science & Eng.
School of Information
University of Michigan
harmank@umich.edu

**Raymond Fok**
Computer Science & Eng.
University of Michigan
rayfok@umich.edu



## Abstract
Situated interactions provide a powerful means of leveraging physical context to make possible rich, efficient interactions between AI agents and human users. However, the use of speech and gestures to make sense of this context is not accessible for people with certain types of motor impairments. We propose a direction of work that aims to combine context from multiple interactional sources with collective human intelligence to help overcome these accessibility challenges.

## Author Keywords
Accessibility, access technology, motor impairments, situated interaction, speech, gesture, crowdsourcing

## Introduction and Related Work
Situated interaction leverages a physical environment's context to make communication richer and more efficient between an AI agent and a human user [3]. These interactions leverage gesture and references to physical surroundings, in addition to speech, to make sense of the interaction context. However, these speech- and gesture-based interactions are not always accessible to people with certain types of motor impairments that may reduce their ability to accurately reference an object via gesture, or may result in modified speech patterns. As a simple hypothetical example, imagine a person asks an

assistive robot to retrieve an object using the natural language (NL) request: "Bring me that adapter." Figure 2 shows a general setup of situated interactions: *speech*, *language*, *perception*, and *gesture*. Speech processing allows us to understand and transcribe the language, and language processing helps us understand a described reference to an object (e.g., "adapter"). Perception lets the machine understand what objects are in the environment, and gesture recognition allows the system to understand physical references made by the user (e.g., user points directly at the adapter).

Automated methods cannot solve any of these problems perfectly. Issues in perception and in natural language understanding are common for all users (e.g., the AI agent might not know the object being referred to by the user, or the user might not use the right language to refer to the object). Worse, some of these problems are exacerbated for people with motor impairments. For instance, modified speech patterns make transcription from speech significantly more difficult [1].

Prior work in crowd-powered systems has tried to bridge some of these accessibility gaps [2]. More specifically, prior research has looked at leveraging crowdsourcing for speech [9], browsing websites [11], maintaining user privacy [8], recognizing multimodal gestures [12], and for collaborative language grounding [4], all of which could power new access technologies.

## Challenges

There still remain open challenges in making situated interactions more accessible. While current perception and natural language processing capabilities need advancements at a high level (i.e., to make these better for all users), gesture and speech recognition are areas where issues are compounded for people with motor impairments. We focus on these two components in our proposal.

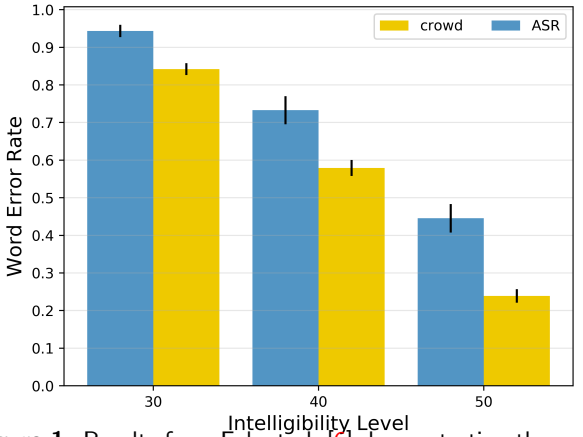


**Figure 1:** Results from Fok et al. [6] demonstrating the difficulty of captioning speakers with modified speech (here, deaf speech) for both automatic speech recognition (ASR) systems and individual human (crowd) captionists. The intelligibility rating of the speech segments is shown on the horizontal axis (higher is more intelligible). As intelligibility falls, error rate quickly rises for both sources of captions.

*Gestures*

For many motor-impaired users, it is difficult to accurately gesture and point to the objects that they want to interact with. For instance, Mott et al. [10] observed 10 people with motor impairments and find that it is difficult for them to accurately select a single point, since their touch behaviors create multiple contact regions. Accurately pointing to an object is a major challenge, and no existing work has explored how to resolve reference ambiguities of this type.

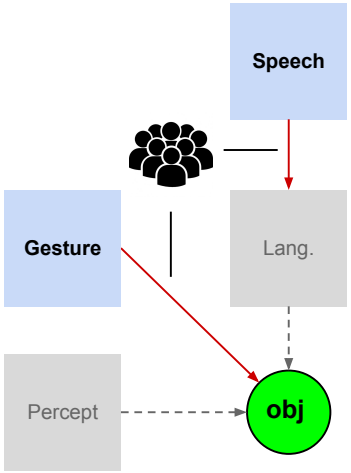


**Figure 2:** General setup of situated interactions. While issues of natural language understanding (Lang.) and perception (Percept) are common, these are the same for motor impaired users as for anyone else, and thus are not our focus here. We believe that crowds provide a powerful and highly available means of addressing challenges in speech and gesture understanding, but new ways to jointly leverage context are needed.

*Speech*
On the other hand, spoken language has been used as a way to refine user gestures. If a user were to point to an object, that pointing gesture can be refined with a spoken "that bottle," which helps to more easily disambiguate which object the user specified. Often, these refinements can come from conversational context clues [5].

Unfortunately, due to modified speech patterns caused by motor impairments in speech-related muscle regions, this refinement approach might not work for such cases, as the assistive robot might not understand the user. Fok et al. [6] show the difficulty in captioning modified speech (Figure 1). As a result, spoken language by itself cannot be relied upon to refine user gestures as reliably as unmodified speech.

## Proposed Approach

To address these two challenges, we propose using collective human intelligence (via crowdsourcing) to bridge some of the key system comprehension gaps we identify above.

*The Importance of Context*
Current crowd-powered approaches use context from pairwise intersections between these four components in Figure 2 to overcome some of these accessibility gaps. For instance, if we know that there is a cup in the scene (perception) and the user says "cup" (speech), the system can use that recognized object to narrow down the list of object candidates (similar to how language models narrow down candidate words).

As seen in Fok et al. [6], though speech transcription via crowds shows promise, it is not enough to have just pairwise intersections for context. With an iterative workflow that let crowd workers build up on prior partial answers, the authors saw up to 74% decrease in word error rate, and a 26% relative improvement when given use-case context.

Furthermore, there remains an additional open challenge for situated interaction references, as no one has looked at "noisy" gestures for referencing objects in physical environments (though as we saw earlier, Mott et al. [10] looked at this for touch interfaces). This object referencing is itself based on the machine's ability to perceive, which has been explored by EURECA [7], but not resolved, and not when the initial request is low-confidence.

*Jointly Leveraging Multiple Sources of Context*
Another open challenge is as follows: How do we combine and leverage context in SI settings across all four components to help workers bridge the two identified gaps? Multiple sources of context can help to resolve these complex references, or to further improve people's ability to disambiguate references. However, there remains the key open question of how to present this joint context while avoiding an increase in people's cognitive load.

## Conclusions and Future Work

By leveraging a physical environment's context, we can make more efficient and richer interactions between AI agents and human users possible. However, there remain open challenges in order to make these interactions more accessible to people with motor impairments.

Future work may investigate how to leverage context between speech, language, references, gestures, and machine perception to best bridge accessibilty gaps. More specifically, how can we design systems and approaches to *share* context between these various facets of interaction?

## Acknowledgments

The authors would like to thank Martez Mott and Skanda Palani for their contributions to this work.

## References

[1] Bigham, J. P., Kushalnagar, R., Huang, T.-H. K., Flores, J. P., and Savage, S. On how deaf people might use speech to control devices. In *Proc. of ACM SIGACCESS Conference on Computers and Accessibility*, ACM (2017), 383–384.

[2] Bigham, J. P., Ladner, R. E., and Borodin, Y. The design of human-powered access technology. In *Proc. of ASSETS*, ACM (2011), 3–10.

[3] Bohus, D., and Horvitz, E. Models for multiparty engagement in open-world dialog. In *Proc. of SIGDIAL*, Association for Computational Linguistics (2009), 225–234.

[4] Chai, J. Y., Fang, R., Liu, C., and She, L. Collaborative language grounding toward situated human-robot dialogue. *AI Magazine 37*, 4 (2016).

[5] Fang, R., Doering, M., and Chai, J. Y. Embodied collaborative referring expression generation in situated human-robot interaction. In *Proc. of HRI*, ACM (2015), 271–278.

[6] Fok, R., Kaur, H., Palani, S., Mott, M. E., and Lasecki, W. S. Towards more robust speech interactions for deaf and hard of hearing users.

[7] Gouravajhala, S. R., Yim, J., Desingh, K., Huang, Y., Jenkins, O. C., and Lasecki, W. S. Eureca: Enhanced understanding of real environments via crowd assistance.

[8] Kaur, H., Gordon, M., Yang, Y., Bigham, J. P., Teevan, J., Kamar, E., and Lasecki, W. S. Crowdmask: Using crowds to preserve privacy in crowd-powered systems via progressive filtering. In *Proc. of HCOMP*, vol. 17 (2017).

[9] Lasecki, W., Miller, C., Sadilek, A., Abumoussa, A., Borrello, D., Kushalnagar, R., and Bigham, J. Real-time captioning by groups of non-experts. In *Proc. of ACM User Interface Software and Technology (UIST)*, ACM (2012), 23–34.

[10] Mott, M. E., Vatavu, R.-D., Kane, S. K., and Wobbrock, J. O. Smart touch: Improving touch accuracy for people with motor impairments with template matching. In *Proc. of CHI*, ACM (2016), 1934–1946.

[11] Oney, S., Lundgard, A., Krosnick, R., Nebeling, M., and Lasecki, W. S. Arboretum and arbility: Improving web accessibility through a shared browsing architecture. In *Proc. of UIST* (2018).

[12] Speicher, M., and Nebeling, M. Gesturewiz: A human-powered gesture design environment for user interface prototypes. In *Proc. of CHI*, ACM (2018), 107.