Talk-To-Me: Designing Speech Input for Public Spaces

Giovanna Nunes Vilaza
UCL Interaction Centre
University College London
London, UK, WC1E 6BT
giovanna.vilaza.16@ucl.ac.uk

Danilo Di Cuia
UCL Interaction Centre
University College London
London, UK, WC1E 6BT
danilo.cuia.15@ucl.ac.uk

Yvonne Rogers
UCL Interaction Centre
University College London
London, UK, WC1E 6BT
y.rogers@ucl.ac.uk

CHI'18 Workshop on (Un)Acceptable!?!—Re-thinking the Social Acceptability of Emerging Technologies, April 21, 2018, Montreal, QC, Canada.

Abstract

The challenges of deploying interactive technology in public spaces are well known by academia and industry. Even though much effort has been put in designing public interactions that rely on gestures, typing and tangible buttons, speech recognition is not often the choice. This may not be a surprise, considering how uncomfortable passersby might feel to be seen talking to a machine, and the frustration felt when their input is not correctly recognized. Despite this resistance, speech input can be highly desirable as a way to collect open-ended answers. Therefore, a physical prototype was designed to investigate how speech recognition could be used to foster indirect communication between people in the same public space. However, during pilot studies, concerns about social acceptability raised interesting points for further discussions.

Author Keywords

Public devices; speech recognition; technology acceptance; in-the-wild studies.

ACM Classification Keywords

H.5.m [Information interfaces and presentation (e.g., HCI)]: Miscellaneous

Introduction

Ideally, speech is a natural and straightforward way of providing input to computers. Such hands-free control does not require previous practice, and in principle, can be faster and more expressive than typing. It is also a safer modality when users need to use their hands to perform other tasks, like driving [2] or surgeries [7].

Even though it has advantages, speech recognition is not simple, as it has to deal with ambiguities, semantics, the variability of speakers, pronunciation and intonation [8]. In addition, voice recording in public spaces can be tricky, because the presence of an audience can discourage passersby to interact for fear of being judged [9]. Also, different contexts lead to different norms of what is socially acceptable [1]. Therefore, adequate design decisions need to be considered before choosing when and where to use speech recognition.

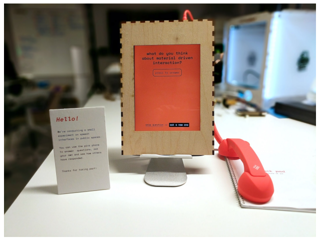


Figure 1: Talk-To-Me physical prototype: telephone and tablet

Given such challenges, a prototype was created to investigate what might be required for speech to be accepted in public spaces. When the input modality is made of spoken words, it is expected that users will be more self-conscious than when they discreetly type their answers. Also, their motivation to interact might be related to what they can gain with it. This paper describes and discusses some of the insights we got from pilot studies.

Related Work

Back in 2004, it was argued that speech input in public spaces was uncommon because this feature was not present in home computers and users could potentially develop unrealistic expectations of the capabilities of the device[4]. However, speech input has seen an increased acceptability of voice-control on mobile phones and home assistants in the last years [5].

When it comes to speech input for public spaces, the OK-net was one of the first instances. It was a public kiosk that allowed speech input for a more natural way to perform queries [11]. However, these queries had to be performed using a quite limited set of commands which the machine was programmed to recognize, constraining its utility. Furthermore, whenever the commands were not recognized, the interaction took more time than typing, contradicting the expectations of a more natural and straightforward input.

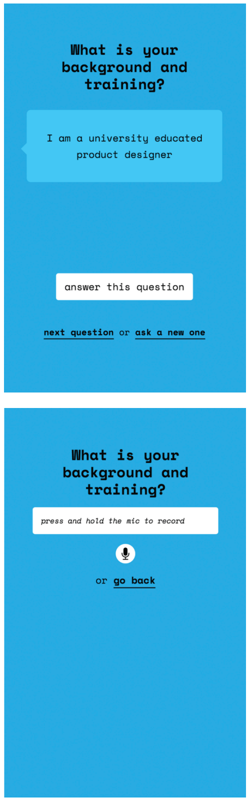
Additionally, conveying the speech modality can be tricky in public. In another study with intelligent kiosks, passersby felt apprehensive of touching the display, and they did not understand they were supposed to speak with it [6]. The lack of physical affordances such as a close-talk headset probably did not help with conveying that speech was the input modality. The microphone also picked up background noise, which made it difficult for the voice to be accurately recorded.

A more recent development was VoiceYourView, designed for recording and displaying open feedbacks about a public library [12]. A telephone was used as a microphone, and technology at that moment already allowed real-time transcription of the voice input, which opened space for more elaborated replies rather than just commands. This turned out to be a quite successful idea, especially because library visitors were keen to leave their opinions. However, its acceptability was questioned by the users, who still felt uneasy to speak to a machine. Interviewees said they were not comfortable with being observed by others, some elderly felt intimidated by the technology, whilst some adults thought the device looked like a toy.

Prototype

Building on previous work, a physical prototype was designed: Talk-To-Me. Its purpose was to allow a group of people at the same public space to ask and answer questions that they created themselves, as a way to get them to know each other through time. The idea was that those who created a question would feel curious to see what others had answered afterwards, and it would also be an entertaining gadget to have at an office space or event, for example.

A bright orange telephone was used as a microphone to record the answers, and it was attached to a tablet (see Figure 1). The tablet was running a web app that used the Chrome Speech API to transcribe the recorded answers in real-time and the text was displayed on the screen of the tablet (see Figure 2). As users reply to questions, their answers are stored and displayed on the tablet, so that other users can see what people have said so far. Given its goal to be placed in public spaces, the physical design and the choice for a striking telephone was meant to be attractive and to easily convey interactiveness.



What is your background and training?

I am a university educated product designer

answer this question
next question or ask a new one
What is your background and training?
press and hold the mic to record
or go back

Figure 2: Talk-To-Me interface for recording answers

Pilot studies

An initial set of eight questions was created by the research team, targeted at the other researchers in the building. They were short questions (between 39 and 74 characters) about their plans for the holidays, their research interests and opinions about the office space. Usability tests were then conducted in the lab: five think-aloud studies were followed by short semi-structured interviews.

Some acceptability concerns already started to appear. One participant said: *"I think it is recording even when I am not answering, it feels weird"*. Another one believed that their voice was being recorded and could be later on be used against them. This indicated a latent issue with privacy and the fear of the device being used for other purposes not disclosed. In addition, there were times when the speech recognition was not working properly, which left participants feeling quite frustrated. This was pointed out as a big drawback: *"The algorithm that detects speech does not work well, I think I would give up really quickly because of that"*.

Furthermore, there were issues with finding a good context for deployment. At first, the prototype was placed at two office spaces for an hour each (kitchen and entrance hall). The same set of questions of the usability tests was used. Even though people noticed the device, they were not approaching it. Short interviews showed that they did not want to be overheard by their colleagues. They also did not want to be seen performing an action that contradicts their expected role at work. When placed for an hour at a coffee shop at the university, with questions about the holiday season, no one was seen approaching the device as well.

In another pilot, the device was placed in a small event for urban planners. The research team pre-loaded three short questions, about the theme of the talk and people's expectations. During one hour, four users were observed interacting with the prototype and they recorded answers and questions spontaneously. All of them answered the first question which was "What do you expect to get from the event?". They said *"networking"*, *"learn more about the topic"*, *"meet people in health and city space"* and *"Seymour diamond design"* (probably a recognition error). Two of them created new questions, which were: "What effects does your city have on your mental health?" and "What is your background and training?".

Whenever a user answered a question, the system displayed it immediately, next to the other answers, but there was no information about how the recorded data was going to be used. During the semi-structured interviews, one participant explained that it was normal to use speech to give commands to the phone and to a home assistant device. However, the experience of doing that in public was slightly less comfortable for this user, especially when the room is quiet and people are overhearing everything.

Discussions

The current trend of voice-control on mobile phones and home assistants is probably changing the perception people have about speech input interfaces. However, applying this modality to public spaces might not be straightforward. The findings presented here are preliminary but they already point to some potential sources of unacceptability.

First, it could be that the purpose of the device was not appealing enough to make people stop their current activities and fully engage with it. A project like VoiceYourView was successful in getting users to submit feedback about a space they frequently go [12]. Further tests are required to understand to what extent the purpose of the installation can hinder engagement. Could it be that people are not interested in answering and asking questions between each other?

When the device was deployed in a space where people knew each other, users were concerned about being overheard and judged. At the coffee shop, people avoided approaching it. On the other hand, in an event full of strangers, they were more open to it. They also easily captured the purpose of the prototype, as they created questions that were relevant to the other attendees. Whilst the role of context has been studied for public installations [1], how to predict which contexts will be more conducive for people to speak up?

When it comes to the technical issues, they can lead to a significant drop in engagement. In a home environment, users might need to repeat a command several times, but if they really want to get the machine to perform an action, they will do it anyway [5]. However, in a public installation, passersby might not be bothered to speak multiple times, especially if they are busy with their own activities [3]. How can we keep users engaged even during the occasional system faults?

Moreover, the prototype functioned like a *recording machine*, as it did not speak back to the user. It could be that making the interaction more similar to a dialogue would have increased its acceptability. The interaction with voice-controlled home assistants resembles more a conversation [5] as well as in some virtual guides [10]. Could it be that conversational style leads to a more natural interaction? Would that help to decrease the feeling of being *spied*?

Some considerations to help mitigating acceptance issues include adding a more playful task to spark the users' interest. Allowing multiple people to play at the same time could make the situation appear less frightening. In occasions such as informal gatherings, events, group meetings, a more playful behaviour is allowed and expected, which can make users feel more at ease.

Finally, even though speech might not be as discreet as other modalities, it should not be simply avoided. The cases presented indicate that people might feel embarrassed and concerned with speaking in front of others. However, through the discussions of these questions, and future iterations on the prototype, we can better understand where, when and how speech input should be used.

REFERENCES

1. Imeh Akpan, Paul Marshall, Jon Bird, and Daniel Harrison. 2013. Exploring the effects of space and place on engagement with an interactive installation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*. ACM Press, New York, New York, USA, 2213.

2. S. W. Hamerich. 2007. Towards advanced speech driven navigation systems for cars. In *2007 3rd IET International Conference on Intelligent Environments*. 247–250.

3. Elaine M Huang, Anna Koster, and Jan Borchers. 2008. LNCS 5013 - Overcoming Assumptions and Uncovering Practices: When Does the Public Really Look at Public Displays? *LNCS* 5013 (2008), 228–243.

4. Masaki Ida, Hiroyuki Mori, Satoshi Nakamura, and Kiyohiro Shikano. 2004. A noise-robust speech input interface for information kiosk terminals. *Electronics and Communications in Japan (Part II: Electronics)* 87, 12 (2004), 51–61.

5. Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf Between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 5286–5297.

6. Lee McCauley and Sidney D'Mello. 2006. MIKI: a speech enabled intelligent kiosk. In *Intelligent Virtual Agents*. Springer, 132–144.

7. Helena M. Mentis, Kenton O'Hara, Gerardo Gonzalez, Abigail Sellen, Robert Corish, Antonio Criminisi, Rikin Trivedi, and Pierre Theodore. 2015. Voice or Gesture in the Operating Room. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '15)*. ACM, New York, NY, USA, 773–780.

8. Cosmin Munteanu and Gerald Penn. 2015. Speech-based Interaction: Myths, Challenges, and Opportunities. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '15)*. ACM, New York, NY, USA, 2483–2484.

9. Kenton O'Hara, Maxine Glancy, and Simon Robertshaw. 2008. Understanding Collective Play in an Urban Screen Game. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work (CSCW '08)*. ACM, New York, NY, USA, 67–76.

10. David Traum, Priti Aggarwal, Ron Artstein, Susan Foutz, Jillian Gerten, Athanasios Katsamanis, Anton Leuski, Dan Noren, and William Swartout. 2012. Ada and Grace: Direct Interaction with Museum Visitors. In *Intelligent Virtual Agents*, Yukiko Nakano, Michael Neff, Ana Paiva, and Marilyn Walker (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 245–251.

11. Max Van Kleek, Buddhika Kottahachchi, Tyler Horton, and Paul Cavallaro. 2004. Designing speech interfaces for kiosks. In *Student Oxygen Workshop Proceedings*.

12. Jon Whittle, William Simm, Maria-Angela Ferrario, Katerina Frankova, Laurence Garton, Andrée Woodcock, Jane Binner, Aom Ariyatum, and others. 2010. VoiceYourView: collecting real-time feedback on the design of public spaces. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*. ACM, 41–50.