Hey Computer, Can We Hit the Reset Button on Speech?

Benett Axtell
Christine Murad
University of Toronto, TAGlab
Toronto, ON, Canada
benett@taglab.ca
cmurad@taglab.ca

Benjamin R. Cowan
School of Information & Communication Studies, University College Dublin
Dublin, Ireland
benjamin.cowan@ucd.ie

Cosmin Munteanu
University of Toronto, TAGlab and University of Toronto Mississauga, ICCIT
Toronto, ON, Canada
cosmin@taglab.ca

Leigh Clark
Philip Doyle
School of Information & Communication Studies, University College Dublin
Dublin, Ireland
leigh.clark@ucd.ie
philip.doyle1@ucdconnect.ie

Abstract
Conversational interfaces are seen as the "next thing". Intelligent Personal Assistants are standard in every smartphone, and Conversational Agents are moving into homes. However, these tools have not advanced far beyond classic examples, such as ELIZA, and are not truly conversational. The desire for these systems to be conversational has become the default in the field, at the expense of other modes for speech interaction. Furthermore, the design of conversational UIs lacks evidence or insight from typical design methods like co-design, despite their prevalence in HCI. Here we explore previous works including users' needs in the design of speech-enabled tools, and reflect on what can be learned. We find that research in this area is stuck on conversational interfaces as the only option. Research on speech interfaces needs to be reset so that the needs of users are included and different speech interactions are explored.

Author Keywords
Speech interaction; Co-design;

ACM Classification Keywords
H.5.2 [User interfaces]: Voice I/O, Natural language, User-centered design, and Evaluation/methodology.

Introduction

Conversational interfaces are quickly becoming ubiquitous in our homes and in our pockets. Voice-based interactions are seen as the "next thing" in industry, with much media hype being attached to the emerging tools and applications (e.g., [23]). The optimal interaction paradigm for speech is consistently characterized as conversation-based, with much effort being used to develop conversational speech interfaces [2]. As things stand, ther are two areas lacking attention: first, assessment of the conversation as a metaphor for speech interaction, and second the relevance of conversation to particular use cases. We argue that this focus on conversational interaction has caused core HCI principles to be excluded from the development of speech interfaces, and that research for speech interaction needs to be reset.

This paper proposes that there are many options for speech interaction beyond the current conversational question/answer interaction, and that designing with users is missing in the development of useful human-centred speech interactions. Many of these options have not been fully embraced because conversational interactions dominate both research and industry. Our paper explores these types of interactions, questioning the need for conversation as a metaphor.

Current Voice Interactions are not Conversational

Commercially advertised "conversational" interfaces are currently far from conversational. In reality, simple question/answer routines are the norm [14]. Conversational Agents (CAs – including Intelligent Personal Assistants - IPAs) like Google Home and Amazon Echo employ command-based interaction, rarely including functionality that would be required for a realistic dialog (e.g., saving the context of previous commands, developing common ground during dialog, structuring responses using knowledge about conversational turn-taking and dynamics). Yet users perceive these systems to have far more human-like conversational abilities than is currently the case [2,10,15,17]. In reality, rather than being a "natural" user interface, these interactions tend to be learned through trial and error, sometimes guided by written instructions. This has not moved much farther from the interaction capabilities of ELIZA [25].

That said, much effort has focused on developing the artificial intelligence of these devices to make the experience more human-like and conversational. An example of this is Shabette Concier, a voice-based agent that responds to non-structured phrases of speech, rather than being constrained to command-based interaction [24]. Bowden suggests using data based on actual human conversations in order to drive models of language so as to make voice-based interaction more akin to actual human-human interaction [4]. Similar approaches to this are seen in prominent work in the speech technology field [11]. Beyond the technical issues of processing, understanding, and responding to natural speech, the approach of designing human-like conversational agents might be inappropriate [2,10,15,17]. We argue that the idea of the human-like, conversational personal assistant seems to have become the gold standard for speech interaction. This gold standard view needs to be reset.

Do We Need Conversation?

The reality is, a mimicry of human-like conversation, with the required advancements in speech recognition and understanding, are not required for useful voice-based interfaces [2,17]. In fact, for a number of interaction scenarios, attempting to mimic human-human conversation may not be a practical approach [22]. As many speech devices (e.g. Amazon Echo, Google Home) lack a graphical interface, information can only be presented through audio [7], leading to slow interaction and increased cognitive effort [26], which may be exacerbated through using human-like conversation. These barriers mean that CAs tend to take on question/answer dialog structures [14], but because of this, they are dismissed as non-functional. This dismissal by users is partly due to the misalignment of expectations versus actual interaction experience [2,8,10,15].

Rather than driving towards conversational interaction, designs should instead explore using alternate modalities that complement and support the user's interaction. Though speech interaction has often been viewed as a user speaking directly to an interface, there are many varied and effective speech-based applications that do not involve having a direct conversation with a speech agent [17]. As seen across HCI research, tailoring the type of interaction to the task creates well-designed experiences, and designing with speech is no different [2,22]. If we look beyond the glamour of conversational voice interaction, we can see that there are many different modalities for voice that should be considered in the endeavour to build usable voice-based interfaces.

Non-Conversational Speech Applications

As stated, the focus on supporting conversational interactions is not well suited to the realities of speech as a modality. This is especially true considering that speech is a more usable modality in multimodal settings than on its own [17,22]. When tasks are more complex and present higher cognitive demands, users demonstrate an implicit preference for interfaces that afford simultaneous use of multiple input modalities (e.g., speech and touch) [19]. Even in high-risk situations, such as designing for military training programs, errors in speech recognition can be overcome with multimodal interactions. For example, Fournier et al use field observations and feedback from course instructors to design multimodal interactions that support low-quality speech recognition [13].

Despite the continued focus on IPAs and CAs, there are several examples of diverse, usable speech interactions [e.g., 16]. Some applications only require brief speech interaction, like question/answer or computer-prompted speech, and do not need to be wrapped in a conversation. The Web on the Wall project explored user-elicited gesture and/or spoken commands for web browsing on a large, shared screen, and found that speech was more commonly chosen, since it was seen as an easy modality for simple browser commands [16]. The ALEX project supports low-literate users with, among other tools, the ability to practice their pronunciation by mimicking a prompt spoken by the tablet [18]. These interactions, designed with user input, keep the speech interaction very simple and targeted to only relevant activities. The nature of these tasks is not driven by a dialog, so there is no need for or attempt at conversation.

A different use of speech as a modality is implicit interactions, such as reading aloud, reminiscing, or brainstorming. These interactions are often part of a larger task and interfaces can build upon this casual speech. Frame of Mind, a tool for digital picture interactions designed using contextual inquiries, organizes digital picture collections based on the memories shared between family members [1]. The ALLT e-reader for readers with low vision and their families, helps casual readers to read-aloud, or read along with a recording, using simple highlighting of their current sentence or word [12]. Both these tools take advantage of the speech implicit to the given activities to create a richer experience. These interactions are not directed at the device, and human-computer conversation would only interrupt their process.

Thinking About Design is Critical

None of the speech interactions presented above have true conversation at their core. They were created for distinct tasks and designed by and with users. The resulting speech interactions show a broad range of expectations that users have for this space, from pronunciation guidance to ongoing storytelling, as well as the potential for the future of speech interaction. It is critical that we consider what users want speech to do as well as what benefits speech bring to interaction in specific contexts and tasks.

As well as considering where speech can benefit interaction, significant thought is needed as to the metaphor for interaction that is designed into speech devices. Work has shown that users model interactions with computers based on heuristic models of interactions with other humans [9]. Yet there are clear differences in the abilities of humans and computers to be competent conversational partners [5]. These differences should be driving design decisions. Unfortunately, there is currently little research aimed at understanding how design impacts user perceptions of those differences. Perhaps in light of this, most speech interfaces are designed based on a basic and simplistic mapping between human-human and human-computer dialog, consistently aiming to signal humanness to scaffold user models of interaction.

To address these issues, we must embrace design methods that place the user at the centre of speech interface development. Co-design and human-centred methods are commonly used in HCI to understand what designs might fit a specific task by including user input at all stages of design. Methods such as contextual inquiry [3] and participatory design [21], though common in HCI, are under-used for speech interaction. Recent work has begun to volunteer design recommendations for mobile voice interfaces [7] and explore user experiences and issues with IPAs [20] so as to inform design, but wider exploratory or design-based research is lacking [6]. This means design considerations and heuristics for developing user-centred speech interactions are rare. Given the unprecedented prevalence of these systems, this knowledge is needed to drive future developments of voice interfaces without simply transposing those developed for other modalities [7].

We propose that, in order to become less stuck in the current state of conversational interaction, design research for speech interaction should be reset. We need to build a better understanding of expectations for both human-human and human-computer dialog. If we

take a step back, include users in design, and consider diverse types of speech, we can create meaningful, usable voice-based interactions.

Conclusion

Speech interactions could be as diverse as any other modality, but the focus on conversational interactions limits further exploration. Conversational interfaces, like IPAs and CAs, are now ubiquitous, and speech is being added to almost every new technology. However, these are not usable interfaces and are not conversational, which limits adoption. As these tools spread, alternative design research for usable speech interfaces is presenting a variety of interactions, made with user input and an understanding of users' expectations. There are fewer examples to be cited here than we would like, caused by the prevalence of conversational interfaces. This focus is keeping speech interaction design stuck in the pursuit for human-like interaction and perfect understanding of speech, and is holding back potential new and diverse interactions.

Design research for speech interaction needs to hit the reset button. Let the field take a step back, see the possibilities for speech beyond conversational interaction, and find the different and new interactions that are waiting.

References

1. Axtell, B., & Munteanu, C. 2018. Frame of Mind: Using Storytelling for Speech-Based Clustering of Family Pictures. In *Proc. of IUI '18 Companion*.
2. Aylett, M. P., Kristensson, P. O., Whittaker, S., & Vazquez-Alvarez, Y. 2014. None of a CHInd: relationship counselling for HCI and speech technology. In *Proc. of SIGCHI '14 Extended Abstracts*, 749–760.
3. Beyer, H., & Holtzblatt, K. 1997. Principles of Contextual Inquiry. In *Contextual design: defining customer-centered systems*. Elsevier, 41–78.
4. Bowden, K. K., Oraby, S., Misra, A., Wu, J., & Lukin, S. 2017. Data-Driven Dialogue Systems for Social Agents. 1–4.
5. Branigan, H. P., Pickering, M. J., Pearson, J., & McLean, J. F. 2010. Linguistic alignment between people and computers. *Journal of Pragmatics* 42, 9: 2355–2368.
6. Clark, L., Cowan, B., Doyle, P., Garaiadle, D., Gilmartin, E., Aylett, M., Edlund, J., Schloegl, S., Munteanu, C., & Cabral, J. Under Review. The State of Speech in HCI; Trends, Themes and Challenges.
7. Corbett, E., & Weber, A. 2016. What can I say? Addressing user experience challenges of a mobile voice user interface for accessibility. *Proc. of MobileHCI '16*: 72–82.
8. Cowan, B. R. 2014. Understanding speech and language interactions in HCI: The importance of theory-based human-human dialogue research. In *Designing speech and language interactions workshop, SIGCHI '14*.
9. Cowan, B. R., Branigan, H., Begum, H., McKenna, L., & Szekely, E. 2017. They know as much as we do: Knowledge estimation and partner modelling of artificial partners. In *Proc. of CogSci '17*.
10. Cowan, B. R., Pantidi, N., Coyle, D., Morrissey, K., Clarke, P., Al-Shehri, S., Earley, D., & Bandeira, N. 2017. "What Can I Help You with?": Infrequent Users' Experiences of Intelligent Personal Assistants. In *Proc. of MobileHCI '17*, 43:1–43:12.

11. Edlund, J., Gustafson, J., Heldner, M., & Hjalmarsson, A. 2008. Towards human-like spoken dialogue systems. *Speech communication* 50, 8–9: 630–645.
12. Epp, C. D., Munteanu, C., Axtell, B., Ravinthiran, K., Aly, Y., & Mansimov, E. 2017. Finger tracking: facilitating non-commercial content production for mobile e-reading applications. In *Proc. of MobileHCI '17*, 34.
13. Fournier, H., Lapointe, J. F., Kondratova, I., Emond, B., & Munteanu, C. 2012. Crossing the barrier: a scalable simulator for course of fire training. In *I/ITSEC '12*.
14. Gilmartin, E., Cowan, B. R., Vogel, C., & Campbell, N. 2017. Exploring Multiparty Casual Talk for Social Human-Machine Dialogue. In *Speech and Computer* (Lecture Notes in Computer Science), 370–378.
15. Luger, E., & Sellen, A. 2016. Like having a really bad PA: the gulf between user expectation and experience of conversational agents. In *Proc. of SIGCHI '16*, 5286–5297.
16. Morris, M. R. 2012. Web on the Wall: Insights from a Multimodal Interaction Elicitation Study. In *Proc. of ITS '12*, 95–104.
17. Munteanu, C., Irani, P., Oviatt, S., Aylett, M., Penn, G., Pan, S., Sharma, N., Rudzicz, F., Gomez, R., & Cowan, B. 2017. Designing Speech, Acoustic and Multimodal Interactions. In *Proc. of SIGCHI '17 Extended Abstracts*, 601–608.
18. Munteanu, C., Molyneaux, H., Maitland, J., McDonald, D., Leung, R., Fournier, H., & Lumsden, J. 2014. Hidden in plain sight: low-literacy adults in a developed country overcoming social and educational challenges through mobile learning support tools. *Personal and ubiquitous computing* 18, 6: 1455–1469.
19. Oviatt, S., Coulston, R., & Lunsford, R. 2004. When Do We Interact Multimodally?: Cognitive Load and Multimodal Communication Patterns. In *Proc. pf ICMI '04*, 129–136.
20. Porcheron, M., Fischer, J. E., & Sharples, S. 2017. "Do Animals Have Accents?": Talking with Agents in Multi-Party Conversation. In *CSCW '17*.
21. Schuler, D., & Namioka, A. 1993. *Participatory design: Principles and practices*. CRC Press.
22. Shneiderman, B. 2000. The limits of speech recognition. *Communications of the ACM* 43, 9: 63–65.
23. The hottest thing in technology is your voice - Technology & Science - CBC News.
24. Tsujino, K., Iizuka, S., Nakashima, Y., & Isoda, Y. 2013. Speech Recognition and Spoken Language Understanding for Mobile Personal Assistants: A Case Study of "Shabette Concier." *IEEE MDM '13*: 225–228.
25. Weizenbaum, J. 1966. ELIZA- A computer program for the study of natural language communication between men and machine. *Communications of the ACM* 9: 36–45.
26. Yankelovich, N., Levow, G.-A., & Marx, M. 1995. Designing SpeechActs. *Proc. of SIGCHI '95*: 369–376.