"This Is Not What We Wanted": Designing for Conversation with Voice Interfaces

Stuart Reeves, Martin Porcheron, and Joel Fischer
Mixed Reality Lab, School of Computer Science, University of Nottingham, UK

Design is increasingly said to be about constructing conversations with end users [1]. Advances in underlying voice-related [2] technologies, coupled with the spread of voice-driven agents and dedicated devices such as the Amazon Echo, Google Home, and HomePod, lend weight to the notion of so-called conversational interfaces. In spite of the hyped anticipation of an AI-powered future, however, it is not always clear how the vision of conversation with machines measures up to lived reality, or if it is even relevant to actual design problems.

As decades of speech technology research begin to influence the everyday world, HCI needs to develop two things: first, and at its broadest, a program that integrates and bridges speech technologists with human-centered researchers. Second, we need a mature understanding of how this emerging class of voice-enabled devices and services sits within mundane social environments that are routinely saturated with everyday conversation. One way HCI can do this is by reacquainting itself with how talk is accomplished.

> Two issues need clearing up. The broader conversation about conversation conflates different uses of the word; here, we are talking about the application of conversation in the sense of literal verbal utterances to and around speech-detecting and dialogue-managing technology. We are *not* discussing design approaches that might be styled "conversational" (perhaps the latest metaphor with which to sell design work). Second, we need to recognize that the primary enabling force of voice interfaces' spread resides in significant deep-learning-driven advances that have been made on the *recognition* side of these systems (speech to text in particular). The dialogue side is a different story altogether, and therein lies a major challenge, although from a user's point of view, the technical distinction is meaningless.

Studying Voice Interfaces in Use

We have been tackling this latter aspiration head on, in service to the former. Our recent work [3] has been examining hours of Amazon Echo use from domestic settings. The Echo is a speech-enabled smart speaker from Amazon that uses the Alexa Voice Service. Like other offerings from Google or Apple, the Echo is designed to play music, answer questions, and help with functions such as cooking, calendars, and shopping. The Alexa service itself is also being integrated into familiar household appliances and smart home items (e.g., the AmazonBasics microwave, the Nest Learning Thermostat, and the Nest Hello video doorbell), with Alexa acting as a gateway to a household Internet of Things.

As part of the study, an Echo was deployed in five households for a month at a time along with a custom-built recording device (a Conditional Voice Recorder or CVR; https://github.com/MixedRealityLab/conditional-voice-recorder) that records audio continuously from an embedded conference microphone but retains only the last minute in a

temporary buffer. The CVR operated in parallel with its own speech recognition trained for detecting the wake word (in this case, “Alexa”), meaning we were able to store a minute before and minute after periods of Echo use and thus capture something of the circumstances leading up to and following that use. Members of the participating households could see when the CVR was recording and choose to turn it off with the press of a button.

Here we present a set of short transcribed fragments from our data. We adopt an ethnomethodological conversation-analysis approach [4] concerned with how members of social settings—as lay sociologists—treat one another’s activities as primordially *social* actions. For this article, a critical point is that *talk is action*. Language *does* things. When we talk, we are trying to get something done, and done together.

Our study is not designed as a reflection on Amazon Echo or voice interfaces—there are emerging critiques of voice assistants including discussions around their gendered or biased character [5, 6] connected with concerns of inbuilt bias in the training data they draw upon. Instead, we are interested in delving deeper into how participants in the study encountered and dealt with *interactional trouble*. While troubles are a routine feature of everyday conversation [7], many kinds of trouble encountered by users of voice interfaces are unlikely to entirely disappear as a function of incremental advances in underlying technologies; instead, they often rest upon improving design understanding first. The ways in which troubles are encountered and dealt with turn out to be quite revealing and, we hope, offer opportunities for conceptual development around what it means to design interactions with conversationalists. We explore these troubles in two ways: First, we examine how revealing they are of the social organization and moral order of the everyday home environments that these devices sit within. Second, driven by comparing moments of trouble, we identify alternative concepts to conversation when considering the design of voice interfaces.

> The transcriptions in this article use a standard version of Jefferson notation. We do transcription in this way because here it helps us show not just *what* is said (words), but also a bit more about *how* those things are said. In summary: Pauses in seconds and fractions of a second are indicated in parentheses, e.g., (1.5) is a 1.5 second gap; overlapped talk is indicated by [square brackets]; micropauses that are less than 0.3s are indicated with (.); where there are no gaps between turns, we use = to show this; inaudible bits of speech are shown in empty parentheses ( ); laughter-inflected words are indicated with (h) embedded; emphases in utterances use underlining; elongations of words are shown with colons, like thi:::s; double parentheses indicate other kinds of actions happening, e.g., ((laughter)); arrows ↑ indicate words produced at a higher pitch than surrounding talk; talk that is quieter than normal is indicated °with degree symbols°.

Voice Interfaces Are Embedded in the Moral Order of Everyday Life

Perhaps the most obvious thing we notice about participants’ interactions with Alexa is how they become embedded in the complex yet highly ordered life of the home. The world these interactions are going into is built upon everyday and largely unstated shared understandings about how things normally proceed as well as the concomitant moral organization of those understandings. With our first fragment we will begin to unpack these ideas.

In Fragment 1 (Figure 1), we join Nikos and Isabel. Nikos is hosting a New Year's party and is trying to get the Echo he was given as part of the study to play some suitable music.

01	**Nikos**	Alexa
02		(2.6)
03	**Isabel**	play some New Year's music
04		(1.7)
05	**Alexa**	here's a station for jazz music (.) instrumental jazz.
06		(1.4)
07		((music starts playing))
08		(4.4)
09	**Isabel**	Al(h)exa this is not what we w(h)anted
10		((laughter))
11	**Nikos**	Alexa: (0.8) shut up.
12		(0.8)
13	**Isabel**	H↑E:Yuh (0.5) Alex(h)a (.) Nikos apologises for being so
14		rude
15	**Alexa**	hi there
16		(2.2) ((music is still playing))
17	**Nikos**	Alexa stop (0.7) stop
18		((music stops))

Transcript 1. Fragment 1 of interactions between Isabel, Nikos, and Alexa.

Fragment 1. Nikos and Isabel jointly produce the first instruction to Alexa: to "play some New Year's music." Alexa responds (line 05), and Isabel's negative assessment of this response is that the music is "not what we wanted," further reinforced by her laughter. Now, as competent conversationalists, people work within the complexity of categorization routinely [8]. It is *not* categories of genre or artist or song Isabel is asking for—which tend to work more easily as search keywords—but rather a set of quite disparate songs that are category-bound to a particular temporal event. This sequence thus reveals various socially shared and culturally situated assumptions about what constitutes possible categories that might be employed when instructing Alexa to do something as "straightforward" as playing music.

We also see in Fragment 1 a use of the Echo as a prop for shared jokes, involving utterances ostensibly addressed to the device but doing other things for the social situation. On line 09 Isabel laughingly says "this is not what we wanted," which she addresses notionally *to* Alexa but in doing this deftly provides a joke for co-present others to join in with. Through this the device comes to be embedded conversationally in the routine doings of domestic life (in this case, hosting a party and having fun therein) in ways probably not considered by its designers.

Next, Nikos tries to resolve the problem and stop the music playing with "shut up" (line 11), but Isabel then chides him with a third-person "apology" that again uses something similar to

line 09, with its ironic address to Alexa: "Nikos apologizes for being so rude." It's important not to get confused here, however. Isabel is not somehow apologizing *to* the device but rather offering an analysis of Nikos's behavior that is accountable to a particular normative moral order (specifically, being polite). Thus, what we see in this part of the sequence is an exhibit of the shared, agreed-upon sets of ways of acting against which we are held to account. This order is not somehow suspended when addressing Alexa. What is said *to* the device is necessarily often said *around* others. People are *mundanely accountable* for what they say, even when addressing a voice interface; Isabel's response embeds this.

Voice Interfaces Are About Request and Response, Not Conversation

Calling interactions with voice interfaces conversational is perhaps a confusion. We think this idea can gain nuance with a deeper consideration of the individual components of an interaction, namely the requests to, and particularly responses from, voice interfaces.

We now join a family of four—Susan (mom), Carl (dad), Emma (daughter), and Liam (son)—as they eat an evening meal together as they attend to "failures" of the Echo. The Echo deployed in their house is sited on a sideboard near the dining table. In Fragment 2 (Figure 2) they are attempting to get an Alexa Skill (third-party plug-ins to expand the Echo's capabilities) called Quiz Master (a trivia quiz) to start. They initially call this Skill "family quiz."

```
01 Emma     °can you (.) ask for a normal quiz,°=
02 Susan    =Alexa? (0.7) set us a family quiz.
03          (2.5)
04 Alexa    sorry. (.) I can't find the answer to the question I
05          heard
06          (0.4)
07 Emma     Alexa:? (1.0) Set (0.3) a family quiz
08          (2.3)
09 Alexa    sorry. (.) I don't have the answer to that question.
10          (0.4)
11 Liam     Alexa:? (0.9) ↑please set (0.3) a [family quiz.    ]
12                                            [((laughter))    ]
13          (1.2)
14 Alexa    I wasn't able to understand [the question I heard.]
15 Emma                                 [ ((laughs))           ]
16 Liam     ↑beep
17          (0.9)
18 Carl     ALEXA, (0.7) FAMily quiz.
```

Transcript 2. A family of four attempts to access an Alexa Skill.

Fragment 2. The family eventually gets the Quiz Master game started some time later beyond the end of this fragment. Now, we are *not* particularly interested in the specific design problems that can be located in Alexa Skills. Instead, we're interested here in the design of the Alexa responses more generally and particularly how the family treats and

deals with those responses as troublesome matters for repair. By comparing this with another case, we can subsequently examine a broader issue about how request and response is designed.

Susan's initial request to Alexa is an instruction: "Set us a family quiz." Alexa's response is "I can't find the answer to the question I heard." This response explicitly categorizes Susan's utterance as a *question* rather than an instruction. Does this matter? While Alexa's response is an error message that happens to be wrong in some way, a key problem is that it offers little in the way of next actions. By *next actions* we make a conceptual connection with conversation analysis. Conversation analysis offers strong evidence to suggest that when we talk, we are constantly working out how to make sure that our talk is sequentially organized. By *sequentially organized* we mean that one utterance follows the next, and that present utterances set the stage for how future ones are heard/acted upon. This is what Henry Sacks alludes to with his description of the "machinery" of interaction [9] and its retrospective-prospective character.

So, given this point, what happens next in the sequence? Emma has few places to go with her next turn-at-talk, so she repeats Susan's instruction with a slight variation: "Set a family quiz" (line 07). We see this kind of repetition and variation frequently when users are trying to deal with trouble in use. Alexa then responds with another similar question-categorization, "I don't have the answer to that question" (line 09). Liam attempts another variation that displays his recognition of the situation with its troublesome character and transforms the attempt at a further instruction to Alexa into something amusing: "Please set a family quiz" (line 11). This differs from Fragment 1 slightly in that here Liam embeds a humorous turn of phrase into something designed for a response from the Echo, the evidence of which is the shared silence of 1.3s on line 13 (compare with the absence of anything similar on lines 09–11 of Fragment 1). Finally, there is another similar response from Alexa and another even more pared-down attempt from Carl: "Alexa, family quiz" (lines 14 and 18).

This is an example of collaborative repair by the family. Furthermore, these collaboratively produced, minutely varied repetitions of the request to "set a family quiz" seem to be closely aligned with the repeated unhelpfulness of the responses from Alexa. As a social environment, home life frequently turns on offers of help (both explicit and implicit), which emerge frequently to smooth everyday interactions along [10]. This emerges more broadly as a kind of "politics of control" that is worked out as part of the life of the home [3]. Competent conversationalists routinely perform remedial action to repair emerging misunderstandings between themselves and others [7]. But voice-driven devices seem poorly designed to live in this world of constant "fixing," and as a result it is users of them who are thus seeking to repair various sense-making problems that are encountered.

Next we want to contrast Fragment 2 with an *alternate* way these kinds of designed requests and responses might play. While what happens in Fragment 3 (Figure 3) below is also a "failure," it turns out quite differently for the family. The family is trying a different Alexa Skill, a game of Beat the Intro, which plays just the beginning of a song where players must guess the song or artist name.

```
01 Emma     Alexa? (.) [ (1.0)                    play beat the intro
02 Carl                [is it called beat the intro?
03          (1.9)
04 Alexa    you want to hear a station for b b intro. [(0.4)  right?
05 Emma                                               [ no
06          (1.1)
07 Emma     no (.) I don't Alex(h)uh (0.5) (h)No,
08          (1.3)
09 Alexa    alright.
```

Transcript 3. Emma and Carl try to start a game called Beat the Intro.

Fragment 3. We want to draw attention to the response from Alexa on line 04 and what it lets Emma do next after having instructed Alexa to play Beat the Intro. The device's response here incorporates a transcription of the result of its speech-recognition process, "b b intro." Although it is actually a *mis*transcription by the Echo in this case, the response nevertheless *builds* this transcription in and offers a candidate a next action as a question, the action being to "hear a station" formulated as a question, that is, tagged with "right?" The difference between this sequence and Fragment 2 is that here the response *gives Emma a place to go*, and she makes the next move—"no" (lines 05 and 07). The sequence then draws to a close with Alexa's "alright."

Response design here—intentionally or unintentionally—differs a lot from Fragment 2. Here, the response gives participants the interactional resources to move on sequentially, to do the next action and to progress with what they are trying to get done.

We think the concept of conversational design needs to be revised, specifically by instead talking about sequentially organized moves around request and response. We can summarize this notion in the following way. First, responses from Alexa are treated by participants as resources for further action. So responses like "interesting question" or "I didn't understand the question" offer little purchase for that as a result. Second, it seems important to consider how to explicitly design in those resources and embed them in responses. Third, responses enable certain kinds of possible next moves in the sequence but also shut down others. So it's not necessarily about establishing rapport, personality, or some other abstract idea, but instead concretely thinking about how responses and their design enable progressivity for users.

Conclusion

In 2014 Amazon released a vision video (https://www.youtube.com/watch?v=6V5I8HHFTNQ) promoting the Echo as a device embedded within the lifeworld of a typical U.S. family. The video depicted the family gradually learning the capabilities of the Echo and its immersion into home life—answering questions, helping with cooking and homework, telling the news, joking, playing music, and mediating family disputes. This way of thinking about voice-enabled agents turns on the kinds of stereotypical depictions, limitations, and erasures one might expect from a vision [11]. Although Amazon ultimately took the video down after a stream of parodies were

uploaded (and instead began offering far more muted depictions: https://www.youtube.com/watch?v=sulDcHJzcB4), the vision is still useful to consider because it's not *that* different from the way in which voice interfaces tend to be presented and discussed. Broadly, it represents the familiar confections with which a new technology is often gilded, particularly a sense of naturalness of interaction and seamless inhabitation in its users' lifeworld. Devices like this are pitched as things that will naturally "understand" humans, fit in "seamlessly" to our lives, and come to "inhabit" our social spaces. HCI research also sometimes gets caught up in these visions, but of course they are never accurate, as we have shown in our three fragments. Although our participants became highly sensitive to moments when someone was possibly about to address Alexa, the design of voice interfaces is largely predicated on one-at-a-time type interactions, an aspect that is intimately bound up in the technical construction of speech recognition. This rubs up against the real world's complex yet highly ordered multiactivity settings that in reality are very much the norm. As such, the sheer pervasiveness of the social world and its intrusions on current voice-interface design assumptions remains a serious technical and design challenge. Ultimately, embedding voice interfaces into everyday life takes considerable work on the part of their users. In this sense, HCI needs to take a broad look at what that really means for design: how designs become embedded in everyday talk, its practical and moral organization, and whether it makes sense to consider such interactions with voice interfaces as conversations at all.

Endnotes

1. Hall, E. *Conversational Design*. A Book Apart, 2018.

2. We refer to voice specifically rather than speech because we are interested in the role of all manner of features of sequentially organized utterances, including a wide range of phenomena that tend to be ignored in existing speech recognition systems, such as pauses, prosody, etc. In fact, the use of *speech* in *speech technology* represents an *analysis* of voice, selecting certain vocal features that then "count" as meaningful in the course of interaction.

3. Porcheron, M., Fischer, J.E., Reeves, S., and Sharples, S. Voice interfaces in everyday life. *Proc. of the 2018 ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, 2018; http://dx.doi.org/10.1145/3173574.3174214

4. Sacks, H. *Harvey Sacks: Lectures on Conversation*. Basil Publishing, Oxford, 1992.

5. Hannon, C. Gender and status in voice user interfaces. *Interactions 23*, 3 (May–June 2016), 34–37; https://doi.org/10.1145/2897939

6. Hannon, C. Avoiding bias in robot speech. *Interactions 25*, 5 (Sept.–Oct. 2018), 34–37; https://doi.org/10.1145/3236671

7. Schegloff, E., Jefferson, G., and Sacks, H. The preference for self-correction in the organization of repair in conversation. *Language 53*, 2 (1977), 361–382; DOI: https://doi.org/10.2307/413107

8. Schegloff, E.A. A tutorial on membership categorization. *Journal of Pragmatics 39*, 3 (2007), 462–482.

9. Sacks, S. Notes on methodology. In *Structures of Social Action: Studies in Conversation Analysis*. J. Heritage and J. Maxwell Atkinson, eds. Cambridge Univ. Press, Cambridge, 1984, 2–27.

10. Kendrick, K.H. and Drew, P. Recruitment: Offers, requests, and the organization of assistance in interaction. *Research on Language and Social Interaction 49*, 1 (2015), 1–19; DOI: 10.1080/08351813.2016.1126436

11. Reeves, S., Goulden, M., and Dingwall, R. The future as a design problem. *Design Issues 32*, 3 (Summer 2016); http://dx.doi.org/10.1162/DESI_a_00395