ARTICLE FOR *THE PSYCHOLOGIST*

"This is not what we wanted": Talking with voice assistants

Stuart Reeves and Martin Porcheron

The idea of a 'smart personal assistant' that you can speak to in your home is no longer the stuff of science fiction. Apple Homepod, Google Home, and the Amazon Echo are all vying for this role. They are sold (in their millions) as household helpers that let you perform various tasks naturally by just talking to them, whether that's asking them for information, helping out with the cooking by guiding you through a recipe, putting on some music, doing some shopping, or just telling the time.

If you own one of these devices, though, you'll know that the reality is a bit different. Often they don't seem to hear what we say, and when they do respond, the response often betrays a significant lack of understanding of what we really meant. There are now many videos now available online of inexplicable interactions recorded by owners of these devices. Interaction with them is a little 'messy'.

The field of human-computer interaction (HCI), which has strong historical roots in psychology and its application to computer interfaces, is actively exploring not only the role of these new devices in our home life but also how they might be better designed to take the complexity of conversation into account. As HCI researchers, we think that taking a human-centred approach, by looking at the precise details of how people *actually* use language to get things done, will help us better understand the interactional 'mess' and how to design these systems better. In our research we find that users of voice-based assistants often work *very* hard to integrate them into the social setting and deal with the various problems they encounter in use.

Our group—myself, Martin Porcheron, Joel Fischer and Sarah Sharples—have been doing some empirical work (Porcheron et al., 2018) looking at the Amazon Echo, marketed as a voice-based personal assistant that uses the Alexa Voice Service (one uses the wake word "Alexa" to address it). We did fieldwork by collecting audio recordings from five households each deployed with an Echo for a month, capturing what they said to the device but also the conversations they had before, alongside, and after moments of interaction with the device. Informed by a conversation analysis approach (Sacks, 1992) to make sense of this corpus of hundreds of hours of recorded audio data from the home, we have been developing descriptions of the various methods people use to organise their talk with / around the Echo into a coherent conversation.

Let's take just one example. Nikos and Isabel are at a New Year's party and they are trying to get Alexa to play some suitable music. (Numbers in brackets indicate pauses in seconds and fractions of a second.)

Nikos	Alexa
	(2.6)
Isabel	play some New Year's music
	(1.8)
Alexa	here's a station for jazz music,
	instrumental jazz (1.4) ((music starts playing))
	(4.3)
Isabel	Alexa this is not what we wanted ((laughs))
Nikos	Alexa (1.1) shut up!
Isabel	hey! (0.7) Alexa, Nikos apologises for being so rude
Alexa	hi there
	(3.4) ((music is still playing))
Nikos	Alexa stop stop ((music stops))

We have many such examples (180+) of householders' extended 'conversations' with Alexa. Several observations can be made from this short fragment that illustrate features we repeatedly find in these exchanges.

First we can spot a form of use that is never depicted in the adverts: Nikos addresses Alexa with the wake word "Alexa", but then after a pause, Isabel takes over with her own instruction. It is a form of 'speaker selection' but very different to human conversation (Lerner, 2003).

Nikos	Alexa
	(2.6)
Isabel	play some New Year's music

We see this kind of collaboration (and sometimes 'competition') between users of Alexa frequently in our data. The home is a social environment and offers of help (both explicit and implicit) emerge frequently to smooth things along (see Kendrick and Drew (2015)). There is a politics to the control of the device that is worked out as part of the life of the home (Porcheron et al., 2018).

Having asked for "some New Year's music", Alexa responds.

Alexa	here's a station for jazz music,
	instrumental jazz (1.4) ((music starts playing))
	(4.3)
Isabel	Alexa this is not what we wanted ((laughs))

This response is treated negatively by Isabel. There are three interesting things about this.

Firstly, it turns out that Alexa's response is the result of a speech transcription error (we know this from logs). But the potential mismatch between what was said by Isabel ("New Year's music") and what has been captured by the device is never revealed to users; no hesitancy or uncertainty is displayed in the response from Alexa (e.g., a question format could be employed, "did you want to listen to jazz music?"). Competent conversationalists routinely perform remedial action to repair emerging misunderstandings between themselves and others (Schegloff, Jefferson, and Sacks, 1977). But voice-driven devices seem poorly designed to live in a world

of constant verbal 'fixing' – and as a result it is users of them who are constantly seeking to repair various sense-making problems that are encountered.

The second aspect is about Isabel's negative assessment of Alexa's response and the music being "not what we wanted" (and her laughter). The category "New Year's music" turns on various socially shared (and culturally situated) assumptions about what constitutes relevant music to play; as conversationalists we work with the complexity of categorisation routinely (Schegloff, 2007). It is *not* a genre or artist or song Isabel is asking for (which happen to work readily as search keywords).

Thirdly, Isabel laughingly says "this is not what we wanted" which she addresses notionally *to* Alexa but also deftly acts as a joke for co-present others to join in with. We see frequent uses of the Echo as a prop for shared jokes, often involving utterances ostensibly addressed to the device. The role of the tech as a resource for such things is largely absent from demos or sales pitches for voice interfaces, perhaps because doing irony with the device as a prop might be perceived as undermining for a marketing campaign (since it often turns on making the device look 'stupid').

Something interesting happens next. Nikos tries to stop the music playing with "shut up", but Isabel then chides him with a third-person 'apology' ironically addressing the device.

Nikos	Alexa (1.1) shut up!
Isabel	hey! (0.7) Alexa, Nikos apologises for being so rude

This is another feature we repeatedly see: normative moral order—i.e., the shared, agreed-upon sets of ways of acting against which we are held to account—is not somehow suspended when addressing the voice assistant. What is said *to* the device is necessarily often said *around* others. In other words you are accountable for what you say, even to a computer. The Echo like its counterparts is sold as a device to live in the home but in doing so becomes embedded into the fabric of that home, including the established and expected organisation of social conduct. Thus, conduct designed for the device is nevertheless socially implicated conduct. It's important not to get confused here, however. Isabel is not somehow apologising *to* the device but rather offering an analysis of Nikos's behaviour that is accountable to a particular normative moral order ('being polite').

The final part of this exchange sees Isabel's 'apology' being responded to.

Isabel	hey! (0.7) Alexa, Nikos apologises for being so rude
Alexa	hi there
	(3.4) ((music is still playing))
Nikos	Alexa stop stop

There seems to be little sequential coherence between this response and what Isabel said (or Alexa's prior actions, like playing some jazz). This forms a break in the illusion of what the device is doing. Alexa's 'conversation' with the user is really just set of attempts by the device to fulfil 'commands' that it has likely 'heard'. At best voice devices may have a sense of 'state', connecting one utterance by a user to a prior one, however these are still fairly limited exercises in 'slot-filling' for a set of

possible paths (rather like following a simple recipe). For users, however, there is ongoing context being built up all the time and a rich set of implied meanings (e.g., categorisations) that can be used as resources for 'next moves' in the conversation. For Alexa that tracking of and response to the always-building context is severely impoverished and users must thus work around the limitation all the time. We can see this when Nikos—reformulating his prior command, "Alexa (1.1) shut up!"—utters "Alexa stop stop". Nikos does not treat Alexa's greeting "hi there" as a greeting at all (i.e., there is no corresponding paired greeting from him e.g., "hi Alexa"). Instead he carries on with his command to "stop".

Some concluding remarks. Research into how we talk is catching up with the latest developments in 'conversational' interfaces and personal assistants as they become more widespread in everyday life—both via disciplinary hybrids such as our use of conversation analysis in HCI and in conversation analysis itself beginning to examine the organisation of non-human (and human / non-human) interaction (e.g., see Federico (2013) and Pika et al. (2018)).

Our recent work suggests that while many of these new AI-driven systems are designed to support 'conversations' with people, but the reality of their use is that they tend to display significant difficulty with many routine but deeply critical aspects of talk that have been mostly overlooked by speech technology research (which tends to focus on technologically-driven advances). That said, we nevertheless see users of voice-based interfaces going to significant lengths to repair breaks in interaction, sense-making, and often in the course of doing so, innovating possibly novel conversational forms that research into human language and communication has yet to document fully. Of course, what remains *unclear* (a reminder: our study was limited to one month deployments) is how long people will tolerate such interactional clunkiness and whether this leads either to permanent abandonment of these new voice-based personal assistants or, alternatively, increasingly novel ways of speaking which encompass new forms of device-oriented language—new 'ways of talking' that (much like adapting to a mouse and keyboard) are simply accommodations people must develop to get by. It is these questions besides others that we now seek to address in future work.