Designing for End-User Programming through Voice: Developing Study Methodology

Kate Howland
Department of Informatics
University of Sussex
Brighton, BN1 9QJ, UK

James Jackson
Department of Informatics
University of Sussex
Brighton, BN1 9QJ, UK

Abstract

Voice-based interfaces are increasingly seen as an intuitive interface for smart environment control, but there is currently little support for querying, debugging and customising the rules defining the behaviours of connected smart environments through voice. We are in the early-stages of a research project investigating and protoyping support for end-user programming interactions with voice-based interfaces. We are extending and adapting methodologies from research in end-user programming and natural-language interfaces to allow investigation of 'natural expression' of rules through the design and evaluation of prototypes in real-world contexts. We present data from pilot work in a lab setting with Wizard of Oz prototypes, and discuss how this influenced our planned methodology for upcoming studies in domestic settings.

Author Keywords

End-user programming; smart environments; voice interaction design; conversational interfaces; speech.

ACM Classification Keywords

H.5.2. Information interfaces and presentation: User Interfaces – Theory and Methods; D.2.2 Design Tools and Techniques.

Introduction

In consumer technology, there has been a dramatic rise in voice-based interfaces, particularly those which aim to provide a conversational experience. Amazon Echo/ Alexa and Google Home/ Assistant have made voice interfaces a frontrunner for smart home control, but have so far failed to support editing, debugging and authoring of smart home automation rules through speech. Understanding, configuring and customising the rules that define smart environment behaviours are end-user programming (EUP) activities. Currently, these activities must be done using a separate, screen-based interface, as voice interaction is largely limited to triggering pre-defined behaviours. Automation platforms such as IFTTT allow programming of smart home behaviours through trigger-action rules, but have seen little uptake beyond early adopters and tech-savvy hobbyists. There is a gulf between abstract representations of automated behaviours and the concrete real-world environments in which they play out. For example, a user standing next to a smart lamp wanting to understand or reconfigure the rules for its behaviour must turn their attention from the room to a screen, understand and edit a code-like description, and draw a link between a unique identifier and the object in the room. Supporting these activities through a voice interface, with potential to include gesture and proximity data to support disambiguation, could provide more intuitive ways of understanding and programming smart environments.

Programming using natural language long been a goal in end-user and novice programming research, but has so far fallen short of expectations due to fundamental challenges in reaching alignment in communication between human and system. With voice-based interfaces now widely used in intelligent assistants and bots, there is renewed interest in programming through speech, but we lack foundational research on how best users without a programming background can understand and express rules defining smart environment behaviour.

Gathering data on how end-users users 'naturally' express programmatic rules is a well-established approach in EUP research. However, studies of natural expression of programmatic rules for smart environments are typically carried out using toy scenarios in decontextualised settings, and often limited to written responses to survey questions. This means that there is very little data on natural expression of rules through speech, and no data on how co-speech gesture and contextual elements such as proximity support speech when describing rules. In smart home scenarios, the presence of cameras in sensor-enabled environments makes it feasible for additional contextual information to be used to resolve ambiguities and deictic references (e.g. this, there, that). In addition, it is important to recognize the extent to which 'natural' expression is increasingly influenced by expectations from interaction with existing similar systems. In the context of conversational interfaces, it may be more realistic to focus on language alignment between the system and the user.

In the CONVER-SE project, we are examining the challenges of speech programming for smart environments, and investigating how these could be mitigated in a conversational interface. To carry out this research, we are developing methodology by adapting natural expression studies to include capture

of speech, gesture and proximity in situ. We are also investigating the potential to make use of participatory methods such as bodystorming (in which participants play out interactions with an imagined future system), and Wizard of Oz prototypes (in which some or all functionality is implemented by a human)

Background

Previous research on EUP for smart environments has gathered natural language descriptions of rules using empirical methods including online surveys [1, 2], post-it note instruction tasks [3] and interviews [4]. Existing work has led to some consensus, including trigger-action rules as a simple but powerful format [2, 5], an inclination for users to rely on implicit rather than explicit specification [1, 2] and a tendency for them not to mention specific sensors or devices [1, 2, 4]. These studies have provided important insights into the natural expression of tasks and rules for smart environments, however, context has been largely overlooked in this work, and none of these studies were conducted in real-world scenarios. In addition, natural language descriptions have been collected in isolation from other communicative modes, such as gesture. Given the importance of context for smart environments, it is likely that existing findings only provide a limited picture. For example, the finding that end-users do not make reference to specific sensors or equipment, first reported by Truong et al. [1] and validated by the findings of Dey et al. [4] and Ur et al. [2], may well have been influenced by the lack of real-world context in the studies. Referring to sensors that you know exist in your house would be much more likely than referencing hypothetical sensors in a toy scenario. The importance of real world contexts for smart environment EUP research is beginning to be recognized. For example, a recently published EUP study comparing different notation styles for home automation was carried out in real domestic environments [6], but unfortunately the study design did not allow for examination of contextual referencing, or capture of speech, gesture or proximity data. In advance of conducting studies in real environments, we carried out pilot work to help develop appropriate study methodology.

Pilot study

We carried out a pilot study in a lab setting with 6 participants to explore how different study interventions supported gathering of data that could inform the design of an interface for smart environment end-user programming through voice. The participants were 6 students (3 female) studying humanities subjects, aged 18-45, all of who rated their programming experience as ‘none’ (when asked to choose between ‘none’, ‘some’, ‘intermediate’ or ‘expert’).

The pilot study session lasted for 30 minutes, and involved two distinct activities. In the first activity the researcher demonstrated the functionality of some simple sensors and actuators programmed with specific behaviours. For example, when a red RFID tag was placed on a readers a red light came on, and a proximity sensor was wired to a speaker such that a sound started playing and increased in pitch as an object approached.

In the second activity the participants were asked to set up some rules for interaction in an example scenario using some of the demonstrated sensors and actuators.

Over the course of the session the researcher used a number of different approaches to attempt to capture 'natural expression' of computational rules that describe sensor-enabled smart environment behaviours.

The approaches we explored were:

1. Asking participant to describe behaviour of an existing setup (e.g. proximity sensor connected to a speaker, RFID tags connected to lights)
2. Asking participant to describe to the researchers the rules defining planned future behavior
3. Asking participant to imagine they were speaking to a smart environment controller equipped for audio-visual capture and describe the same rules (with a non-functional camera used as a prop)
4. Modelling a rule description by giving an example of a rule (only used as a last case where the participant was very lost and unable to offer a description using other methods).

The pilot study was recorded using video cameras at each end of the room. The relevant sections of the video recording have been transcribed, including basic notation of co-speech gestures and movements. A first pass of analysis has been carried out using mixed methods (counts and content analysis) to determine which methods are promising to develop for further pilots in real domestic contexts.

Fifty-seven utterances were identified as containing full or partial rule specifications in natural language. 21 were produced during the description of existing behaviours, 13 during discussion with the researcher about possible future interactions, and 23 when imagining giving instructions to a controller about future interactions. The most common trigger word used was 'when', which was used 18 times, with a variation of 'whenever' used 1 time. 'If' was used as a trigger 11 times, and 1 time used to specify a conditional: "When the tag is placed on certain… on this one…show up a light if it is the correct card." 'Once' was used 2 times as a trigger. 'As soon as' was used 4 times as a trigger.

Most utterances were phrased in terms of descriptions of hypothetical situations, rather than instructions to the system. For example: "It will only play if it senses that somebody is close"; "When there's pressure here, that would cause this one to light up." Participants were generally much more comfortable in the world of concrete examples rather than abstract programmatic descriptions. Most struggled to switch from a concrete and descriptive mode of thought to an abstract instructional mode. For some, imagining they were speaking to a controller was helpful in focusing their instructions. For example, one participant moved from describing hypothetical scenarios to giving a rule-based instruction when addressing the prop camera: "If someone says 'feeding', skip to feeding chapter." For others, however, this put them in mind of using the system for immediate control rather than programming future behaviours, for example: "Please turn the sound on"; "Zoom in on that, please."

For one participant who found it very hard to understand what was being asked of her, providing an example rule seemed to be a very effective prompt, allowing her to move towards descriptions such as: "It

responds to touch, and then counts"; "It will only play if it senses that somebody is close." Of course, taking this step means that such descriptions can in no way be said to be the participant's 'natural expression'.

Although we did not set out to specifically investigate the role of gesture in this pilot work, we noted that gesture, deictic expressions and practical demonstration were commonly used in describing system behaviour, particularly when acting out imagined future interactions to describe them to the researcher. For example: "As soon as you come up, and select the RFID tags that you, kind of want <mimes placing tags>, to place in the sensor, the video would detect it, and change the video to the object you have selected."

Conclusion

Our early pilot work has allowed us to investigate the extent to which different study interventions prompt natural language descriptions of programmatic rules for smart environments in participants without a programming background. Empirical data of this sort gathered in real domestic contexts is potentially very useful in designing voice-based interactions that allow participants to understand, debug and change the trigger-action type rules defining smart environment behaviour.

However, analyzing the effects of our interventions (particularly the rare step of explicitly modelling correct rule formations) reminded us of the extent to which 'natural' expression is influenced by expectations from interaction with existing systems and technologies, and conversations with humans about the topic. In our pilot, the conversations with the researcher acted in some cases as an elicitation process by which the researcher drew out the separate parts of the trigger-action rule, and the participant rehearsed their ideas about how to describe interaction rules programmatically. In the context of voice-based interfaces, it may not be helpful to fixate on natural expression, and may be more useful to look at how to support language alignment between the system and the user. There is an inherent gulf between the vague and open specifications given by a human, and the fully-specified clarity required by a system. Although true conversational alignment is unlikely to be achievable with an artificially intelligent agent, understanding how alignment is achieved between human conversational partners when discussing trigger-action rules is likely to be illuminating.

Allowing users to use their own language needs to be measured against the need to potentially provide a new vocabulary to allow users to describe unfamiliar concepts and approaches. Considerations such as these feature in many of the published guidelines on designing for voice[1], although these do not currently consider support for understanding, debugging and changing rules for behaviours.

The next steps for us are to further develop our interventions and pilot the approaches in context. We plan to recruit participants with some level of existing smart home functionality implemented, but will seek householders other than those who setup and implemented the system. Our planned contextual study

[1] https://voiceguidelines.clearleft.com/

procedure has three stages, in which participants are asked to: i) interpret, describe and identify problems with existing rules, ii) suggest rules for modified and new behaviours, and iii) bodystorm interactions with a future voice-based system. At each stage the researcher will give increasingly more specific prompts as far as is necessary to elicit full and unambiguous rule specifications. Interactions will be video recorded to capture speech, accompanying gestures and proximity to relevant objects. We would like to investigate the potential to use conversational analysis in examining the data including verbal, gestural and proxemic interactions, to support an empirically based categorisation of the natural expression of trigger-action rules in situ. We are particularly keep to attend this workshop to discuss the challenges in our endeavor, and contribute input from our previous work in end-user and novice programming, as we suspect many interactions with existing voice interfaces involve behaviours such as debugging that cross into this territory.

Acknowledgements

We thank all the participants in the pilot study. The pilot work was funded by the University Sussex Research Development Fund. The CONVER-SE project is funded by the EPSRC (Grant reference: EP/R013993/1).

References

1. Truong, K.N., E.M. Huang, and G.D. Abowd, *CAMP: A magnetic poetry interface for end-user programming of capture applications for the home*, in *Proc. of Ubiquitous Computing*. 2004, Springer. p. 143-160.
2. Ur, B., et al., *Practical trigger-action programming in the smart home*, in *Proc. of Human Factors in Computing Systems*. 2014, ACM. p. 803-812.
3. Perera, C., S. Aghaee, and A. Blackwell, *Natural Notation for the Domestic Internet of Things.* End-User Development, 2015. **9083**: p. 25-41.
4. Dey, A.K., et al., *iCAP: Interactive prototyping of context-aware applications*, in *Proc. of Pervasive Computing*. 2006, Springer. p. 254-271.
5. Catala, A., et al., *A meta-model for dataflow-based rules in smart environments: Evaluating user comprehension and performance.* Science of Computer Programming, 2013. **78**(10): p. 1930-1950.
6. Brich, J., et al., *Exploring End User Programming Needs in Home Automation.* ACM Transactions on Computer-Human Interaction (TOCHI), 2017. **24**(2): p. 11.