Speaking to Machines. User’s work in Complex Voice Interactional Environments

| Moustafa Zouinar  Orange Labs, Paris, France  moustafa.zouinar@orange.com | Julia Velkovska  Orange Labs, Paris, France  julia.velkovska@orange.com |
| --- | --- |

# ABSTRACT

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

*CHI 2020, April 25–30, 2020, Honolulu, HI, USA.*

© 2020 Copyright is held by the owner/author(s). Publication rights licensed to ACM. ACM ISBN 978-1-4503-6708-0/20/04...$15.00.

DOI: <https://doi.org/10.1145/3313831.XXXXXXX>

\*update the above block and DOI per your rights review confirmation (provided after acceptance)

## Digital Domestic Voice Assistants (DVAs) are now part of our daily lives. The study of their uses in ordinary everyday situations remains however understudied. We present a video-based analysis of interactions with these devices in family contexts (14 hours). The data were collected from several weeks and months of use in 23 households, between 2015 and 2018. We show how these devices imply a more or less extensive extra work that the user has to perform to attain her/his goals and handle interactional troubles. Among the different forms of this work, we found a ‘new’ one, namely managing ‘multi-agent’ situations. We outline design implications from these findings.

## Author Keywords

Domestic Voice Assistant; Conversation Analysis; Video Observation; Human-Machine Interaction; Conversational AI, Multi-Agent Interation.

# CSS Concepts

• **Human-centered computing~Empirical studies in HCI**

# INTRODUCTION

Since 2015, when Amazon released the first commercial version of Amazon Echo, smart speakers or Domestic Voice Assistants (DVAs) – as we will call them in this article - are progressively taking a place in the living rooms, kitchens, bedrooms and bathrooms of a growing number of homes. This growing adoption allows us to observe conversational artificial intelligence in action in ordinary homes for the first time.

Although these devices are attracting more and more attention, with a few exceptions [1, 2, 6, 9, 12], qualitative studies of interactions with DVAs in everyday life contexts are still rare.

This article presents a study of these interactions through a video-ethnographic approach which combines interviews and video data collected in 23 homes between 2015 and 2018.

After a discussion of related work, we look at the ‘extra’ work that users have to perform to interact with DVAs. We conclude with a discussion of some design suggestions that can enhance interaction with DVAs. We propose to expand the scope of design to take into account the ecology of conversational services that are emerging. We suggest that design should go beyond the device, the interface or the interaction with an assistant. We argue that the scope of design should be expanded to include creating *Voice based Interactional Environments (VIE)*, which include devices and “conversational” agents.

# Related work

Since the commercial release of Amazon Echo in 2015, we are witnessing a growing number of studies that explore the use and the social consequences of DVAs.

Drawing on the analysis of online reviews of the Amazon Echo, Purington et al. [11] found that users made an emotional connection to Alexa and that personification (use of person pronouns) seems to play an important role in satisfaction with the device. Bentley et al. [2] examined the use over time of Google Home through a quantitative study (analysis of logs). They found that the main domains used are Music, Information and Automation. Exploration of new domains was however low. Users rarely change their practices over time and the way they interact with the device does not significantly change. The authors also observed some demographic differences in use. For example, 18-24 years old users are less likely to listen to music on the device compared with 25-44 year old ones. They also distinguished four distinct user groups: *Heavier* users who have the highest number of commands, sessions per day, and domains used; two *medium* clusters of users, and *light* users who use the fewest commands and domains. These groups also differ in their style of use and the commands used.

Besides these quantitative studies, another set of research explores how people *really* interact with these devices in their homes. Porcheron et al. [9] analyzed interactions with Amazon Echo from a corpus of audio data. They examined how the Echo is embedded into home activities (e.g. dinner) and showed how the use of the device is socially regulated through what they call “politics of control” which refers to the regulation of who can talk to the device, when, and how. These authors also analyzed the sequential organization of interaction with a DVA and the way participants deal with its responses, particularly in troublesome situations (e.g. delays in response, failures). They found that to get the device work, users perform different “repair” practices (e.g. reformulation, repetition, rephrasing, or alteration of prosody). The authors reject the pervasive idea that interactions with DVAs are conversations insofar as the system’s responses and the interactional rules it follows are not emerging in a situated way but are pre-configured by the designers. Drawing on audio-recordings of interactions with Amazon Echo as well, [1] explored the methods deployed by families to repair communication breakdowns. They found five ‘strategies’: prosodic changes, over articulation, semantic adjustments, increased volume, syntactical adjustments, and repetitions. In the same methodological vein, analyzing video recordings of interactions with this DVA, Lahoual and Fréjus [6] show that users encounter interactional difficulties which disrupt their activities. The authors also examined the practices implemented by users to manage these troubles.

The study presented in this article was designed as early as 2015 with the same goals as these works, that is, to understand the use of DVAs in ordinary homes. While related work provides interesting findings about the use of these devices, little is still known about: 1/how users interact with different ‘conversational’ agents trough the same device (what we will call “multi-agent situations”), Our corpus of video-recording of interactions allows us to analyze this issue.

# METHOD

Carried out in France, our study comprised three phases. The first one, which was exploratory, focused on the uses of the first assistants released in the North American market in 2014-2015 (Amazon Echo and Ivee Sleek). As VDAs were not available and used in France at the time of the study, we decided to explore the usage of these devices in an ‘experimental’ way. So, we recruited three bilingual families that speak English and French and we asked them to test a VDA at home for at least one month. Two families (parents with children) were given Ivee; one family (parents with children) tested Amazon Echo. These families were recruited via acquaintances. The second phase (2016-2017) focused on an experimental DVA prototype developed within Orange labs for research purposes. Seven diverse French speaking families (with children), who were given a prototype, participated in this phase. They were asked to test the device at home during at least one month. The participants were recruited in Orange.

The release in France of Google Home in 2017, then Amazon Echo in 2018, allowed us in the third phase (2017-beginning of 2019) to extend our study to households that bought these devices. Thirteen French participants with different profiles (students, executives, employees, self-employed, and former unemployed executive undergoing retraining) were recruited through ads posted on forums and social networks. Participants were living alone, with a parent, or were in a couple (with or without children). Their ages ranged from about 20 to 50 years old. To our knowledge, this third phase is the first study that investigates the usage of DVAs through audio-video data in contexts where the users have voluntarily purchased the devices. All in all, the study was conducted in 23 homes.

The study was conducted in three steps:

- An initial semi-structured interview that took place at the participants' homes. The objective of this first interview was to learn about the households both in terms of their social composition (the people who live there, their family ties, etc.), participants’ digital practices, and the use of voice assistants in the households that were already equipped with them.

- Video recordings of interactions over periods ranging from one to several months. These recordings were made by the participants using a camera that was lent to them during the initial interview. They were instructed to record as often as possible their interactions with the VDA. We have thus compiled a corpus of about fourteen hours of interaction recordings, which represents about 300 sequences ranging from a few seconds to more than 15 minutes (more rare case).

- A final semi-structured interview. The purpose of this interview was to ask participants about their use and appreciation of the devices they had tested or purchased. Self-confrontations [3] were also conducted with some participants during this interview, particularly in the third phase. They were asked about their uses based on video recordings of their interactions with the assistants. This interview also took place in the participants' homes and, where possible, in the presence of several members of the same household.

To thank them for their participation, participants were paid with gift vouchers after this interview.

All interviews, including self-confrontation interviews, were transcribed and analyzed manually according to different predefined themes or those that emerged during the exchange. These interviews are seen as documentary resources that give access to users’ lived experience with voice assistants. To transcribe (table 1) and analyze video data, we relied on conversational analysis [10]. This analytical framework is particularly well suited to analyze interactions, their structure and sequential organization.

|  |  |
| --- | --- |
| (xx) | Inaudible segment |
| (.) or (1) | time intervals (a gap of approximately one tenth of a second or one second respectively) |
| HELLO | Louder voice |
| wo::rd | Extension of the sound or the syllable it follows |
| [thanks] | Overlapping talk |
| = | Latching utterances |
| ((actions)) | Descriptions of events or actions rather than transcriptions |
| / and \ | Rising and falling intonation respectively |

Table 1. Conventions of transcription.

VDAs’ responses and states are identified by the name of the device (Alexa, Ivee or GH for Google Home).

# INTERACTING WITH DOMESTIC VOICE DIGITAL ASSISTANTS: the User’s WOrk

In his latest book on Artificial Intelligence (AI), Harry Collins points out that the narratives about this field are part of the “*world of artifictional intelligence, available through newspapers, books and films*.” [4] This ‘artifictional’ mode - that is, one that overshadows the limits of AI systems - is actually the way in which AI is presented to the public both by its critics and advocates, each overestimating its real capabilities. Regarding DVAs, companies’ advertising operations fully participate in this ‘artifictional’ world, presenting these devices as ‘assistants’, ‘butlers’, even ‘family member’ which facilitate the daily life of the house, thanks to their ability to interact in a “natural” way and therefore to enable users to use them effortlessly. In advertising videos, these devices appear invisible or transparent, easily fitting into multi-activity situations (e.g. people interact with the assistant while cooking or packing). The conversational intelligence of the system, i.e. its ability to understand users’ utterances and manage the dialogue in a fluid manner, is at the heart of these advertisements. However, it is in fact limited and produces an appearance of natural conversation which is a source of interactional troubles [12]. This ‘artifictional’ discourse obscures what users really do when they talk to DVAs, including in terms of additional actions. We mobilized the notion of ‘user’s work’ [12] to make visible what people do in addition to what is provided in the system’s script, or on the ‘administration’ level of the device (installation, configuration, maintenance). We observed this phenomenon throughout the entire study, since 2015 until 2018. As shown by other studies (e.g. [2, 5, 7, 8]), this can take different forms, such as repetition, reformulation by shortening or expanding the request, modification of the prosody (speaking slowly) or the volume of voice (speaking loudly), or coming close to the object. Our data shows that this way of dealing with troubles is not specific to a DVA; we observed it across the three devices that we studied. The following fragment (excerpt 1) where the user tries to play a song by Katy Perry illustrates this work[[1]](#footnote-1):

*Excerpt 1*



Image 1

1. Tom: Alexia go on to YouTube and look up Katy Perry please

2. Alexa: hmm (.) I can’t find the answer to the question I heard

3. Tom: Alexia (.) go on to You Tube (.) now

4. Alexa: hmm (.) I can’t find the answer to the question I heard

5. Tom: Alexia You Tube

((Tom tries again to play a song by Katy Perry. His wife suggests him to find a specific song “Crazy” remixed by Nôze))

6. Tom: Alexia on the amazon music store please find songs like crazy

7. Alexa: here a sample of crazy by (xx) barkley

8. Tom: oh no no Alexia find (1) find crazy=

9. Radio: =Highly unusual [(xx )

10. Tom: [no no Alexia turn that down turn down Alexia Alexia turn that off ALEXIA



Image 2

From line 1 to line 5, we can see how Tom tries to solve the trouble he encounters by reducing his initial request (line 1). As Alexa does not find the answer to the requests formulated by Tom, his wife suggests him to look for a specific song remixed by another artist (Nôze). But he failed again to be understood by Alexa which starts to play a radio station (lines 6 and 7). He then spontaneously comes very close to the device to ask Alexa to stop the radio (line 10 – image 2). Thanks to the video, we are able to observe how the body can be mobilized as a resource for interaction with DVAs to try to manage interactional troubles.

The interviews reveal that the users are quite aware of the work they have to perform to be understood by the system:

“[...] it's a little constraint to think a little more calmly about the whole sentence that we're going to say when it's a little complicated or when we want a wake-up call on a particular day or every day of the week to a particular music. When you want something a little specific, sometimes it's a little long sentences, and it's true that it requires more concentration, whereas in a human conversation, sometimes it's not disturbing that there's a little white space, or if you want to do the same thing on the phone, you'll launch the application, create a new thing, and you'll have time to think.” Luc

In addition, the formulation of requests is subject to a strong time constraint because the user has to produce her/his request at the right time, when the system is ‘listening’ (otherwise the system does not process the utterance). This respondent highlights the differences in the temporality between interaction with DVAs and human conversation on the one hand, and the use of a graphical interface that offers more flexibility on the other. Actually, the user's work extends beyond the formulation of requests. It encompasses all extra-activities performed by users to make the system work, and to understand troubles, as we can see in the following excerpt, where two adult participants (Emma and her daughter Julie) try collaboratively to make sense of the system’s incorrect response (line 2) and to figure out how they should formulate a request concerning the weather (lines 4 and 5):

*Excerpt 2*

1. Emma: hello Ivee what is the temperature in celsius in New York city/

2. Ivee: currently in Montreuil France it is 17 degrees and mostly cloudy

3. Emma: hello ivee what is the temperature in New York city/

4. Emma: maybe we are supposed to just say New York I don’t know

5. Julie: bah (0.5) what’s certain that New York city doesn’t sound Montreuil France

# Users’ work in multi agents situations

DVAs are gradually becoming an entry point to a growing number of third party applications that are managed by specific conversational agents. For example, a home assistant’s user can call an Uber, order a pizza from Domino's or play games. These services are what it is now called ‘skills’ for Amazon Echo and ‘Actions’ for Google Home. The exploration of our video data revealed a new situation which extend the user’s work and generate usability problems: interaction with third-party conversational agents*.* The challenge for users in this ‘multi-agent’ situation is to know which agent is ‘active’, to identify the beginning and end of the interaction with each one as well as the moments of transition, to know what they can do and how to talk to them. In this kind of situation, interaction with several agents can become complicated as shown by the following fragment during which Pierre interacts with three different conversational agents, Alexa and two third-party ones (AlloCiné[[2]](#footnote-2), Télé-loisirs[[3]](#footnote-3)), without always understanding which agent he is talking to at a given moment. After having checked with AlloCiné movies playing this week (non-reproduced part), he asks Alexa about the TV program for the evening (line 24).

*Excerpt 3*

21. AlloCiné: would you like to discover more films/

22. Pierre: No, it's okay, uh::::=

23. AlloCiné: = Sorry (0.2) I'm not sure

24. Pierre: alexa (1) what's on TV tonight

(1.6)

25. AlloCiné: $sorry (0.2) I'm not sure

(0.3)

26. Pierre: alexa (0.8) what's on TV tonight

(1.8)

27. AlloCiné: mm (0.2) I don't know

28. Pierre : alexa (1.2) open télé-loisirs

29. Télé-Loisirs: welcome to télé-loisirs (0.5) Would you like to listen to tonight's program (0.3) a mood Tv channel (0.3) the news selection or asking me for help/

His implicit request to close the AlloCiné application (line 22) is not taken into account by the agent (line 23), which leads to the failure of his two attempts to get information about the TV program (lines 24 and 26). The user is in a complex situation because he thinks he is talking to Alexa, while it is always AlloCiné which answers him. The problem persists until he finds another way to formulate his request in a third repair attempt using the ‘open + application name’ syntax (line 28). This example illustrates how a simple action as “exiting” an application or skill may engender difficulties and extra-work for the user in a multi-agent context.

The use of third-party applications requires from participants to understand a new situation: from the same device, they interact with different services, with different operating modes, commands and sometimes different voices, creating new usability problems. Acting in multi-agent environments can involve a significant amount of understanding and problem-solving work, contrary to the promise of seamless access to third-party services through these devices.

In this section, we have seen how interaction with VDAs requires extra work in different contexts (finding formulations, speaking at the right time, dealing with troubles, managing multi-agent situations, etc.). This work can be an obstacle to appropriation or even a reason of abandonment when it is too costly in terms of effort. In addition, its effectiveness varies from person to person. For instance, we have observed that young children have great difficulty making themselves understood, and did not have the necessary interactional skills to cope with the troubles they encounter. Instead of being accessible to everybody who can talk and understand language, the voice interface of DVAs may not be ‘universal’.

# Discussion and design implications

Our study revealed that the use of DVAs is far from being intuitive and simple as the common belief suggests. We showed that users perform ‘extra’ work to learn to use the devices and to attain their goals when they interact with them. This work takes different forms: learning to talk to the device, finding how to formulate requests, investigating and repairing interactional troubles, managing the interaction with multiple ‘conversational’ agents. As we have seen, this work is experienced as too costly by some participants, and may make the interaction with DVAs cumbersome. This hinders the usefulness and the usage of these devices.

In the light of this finding, it seems that one way to improve the usage of DVAs and enhance their usefulness is to reduce this work and/or make it less difficult and costly for the users. Some design strategies suggested in the literature go in that direction. For example, Porcheron and al. [9] propose that the responses of DVAs should be seen and designed as *interactional* *resources* that support users and help them to interpret or to understand the behavior of the devices, especially when interactional troubles arise. Connected to this view, our study also reveals that it is important to give users better feedback on failures. Moore [7] suggests improving the ‘conversational’ capacities of these systems by implementing repair mechanisms (e.g. repeat repair: repeat request, hearing check, etc.). Another area of improvement revealed by our study is to facilitate the interaction in multi-agent situations. We have seen that users may be lost in these complex automated speech environments: at certain moments of the interaction, they do not know which agent they are talking to and how to talk to it. This problem shows that we should help users to develop a good understanding of which agent they are talking to, what agent is (still) acting as active interlocutor, how to close the interaction with an agent, and what they can precisely do with each of them. Generally speaking, we propose to shift from the design of ‘isolated’ digital voice agents to that of *Voice based Interactional Environments*. This means that designers should be sensitive to the coordination and orchestration of interaction with different agents and different services.

# ACKNOWLEDGMENTS

We thank all the households that participated to our study.

# REFERENCES

1. Erin Beneteau, Olivia K. Richards, Mingrui Zhang, Julie A. Kientz, Jason Yip, and Alexis Hiniker. 2019. Communication Breakdowns Between Families and Alexa. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (CHI '19). ACM, New York, NY, USA. DOI: https://doi.org/10.1145/3290605.3300473
2. Frank Bentley, Chris Luvogt, Max Silverman, Rushani Wirasinghe, Brooke White, and Danielle Lottridge. 2018. Understanding the Long-Term Use of Smart Speaker Assistants. In *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol*. 2, 3, Article 91 (September 2018). DOI:https://doi.org/10.1145/3264901
3. Béatrice Cahour, Pascal Salembier, Moustapha Zouinar. 2016. Analyzing lived experience of activity. *Le travail humain*, 79, 3, 259-284. DOI : 10.3917/th.793.0259
4. Harry Collins. 2018. *Artifictional Intelligence: Against Humanity's Surrender to Computers*. Polity Press.
5. Jiepu Jiang, Wei Jeng, and Daqing He. 2013. How Do Users Respond to Voice Input Errors?: Lexical and Phonetic Query Reformulation in Voice Search. In Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’13), 143–152. DOI: https://doi.org/10.1145/2484028.2484092
6. Dounia Lahoual and Myriam Fréjus. 2019. When Users Assist the Voice Assistants: From Supervision to Failure Resolution. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (CHI EA '19). ACM, New York, NY, USA. DOI: https://doi.org/10.1145/3290607.3299053
7. Robert Moore. 2018. Repeat Repair & Disengagement with Voice-enabled Conversational Agents. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems (CHI EA '18)*. ACM, New York, NY, USA.
8. Hannah R. M. Pelikan and Mathias Broth. 2016. Why That Nao? : How Humans Adapt to a Conventional Humanoid Robot in Taking Turns-at-Talk. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI ’16)*, 4921–4932. DOI: https://doi.org/10.1145/2858036.2858478
9. Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA. DOI: https://doi.org/10.1145/3173574.3174214
10. George Psathas. 1995. *Conversation Analysis*. Thousand Oaks: Sage
11. Amanda Purington, Jessie G. Taft, Shruti Sannon, Natalya N. Bazarova, and Samuel Hardman Taylor. 2017. "Alexa is my new BFF": Social Roles, User Satisfaction, and Personification of the Amazon Echo. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '17)*. ACM, New York, NY, USA, 2853-2859. DOI: https://doi.org/10.1145/3027063.3053246
12. Julia Velkovska, Moustafa Zouinar. 2018. The illusion of natural conversation : interacting with smart assistants in home settings. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI EA '18). ACM, New York, NY, USA.

1. The name of each participant who appears in the excerpts presented here has been changed. [↑](#footnote-ref-1)
2. Tv magazine [↑](#footnote-ref-2)
3. It is a French service that provides information on upcoming or playing now movies. [↑](#footnote-ref-3)