INTEGRATIVE REVIEW ON VOICE INTERACTION TECHNOLOGY

Luiza Mendes Degraf [1]
Francisco Antonio Pereira Fialho[2]

Abstract: *Natural interaction has gained space in Human-Computer Interaction (HCI) studies because it is intuitive, since it focuses on users' spontaneous actions. Voice, for example, can serve as a natural medium in interacting with media objects, and because of this, many voice interaction models have been created in the last sixty years. The diversity and scale of publications on voice interaction makes the research field increasingly complex, making it necessary to systematically structure the field. The objective of this study is the analysis of scientific publications of the last six years on problems of voice interaction, giving special importance to the virtual assistant Amazon Alexa. The result of the analysis suggests 4 (four) different approaches to voice interaction problems: (1) difficulties and opportunities; (2) intentions and utterances; (3) context; and (4) customization.*

Keywords: Voice Interaction; Alexa; Architecture; Problems; Systematic review of literature.

Resumo: A interação natural tem ganhado espaço nos estudos de Interação Humano-Computador (IHC) por ser intuitiva, pois foca nas ações espontâneas dos usuários. A voz, por exemplo, pode servir como meio natural na interação com objetos de mídia e muitos modelos de interação por voz foram criados nos últimos sessenta anos. A diversidade e a escala das publicações sobre interação por voz tornam o campo de pesquisa cada vez mais complexo, tornando necessária uma estruturação sistemática do campo. O objetivo deste estudo é a análise de publicações científicas dos últimos seis anos sobre problemas de interação por voz, dando especial importância à assistente virtual Amazon Alexa. O resultado da análise sugere 4 (quatro) abordagens diferentes para problemas de interação vocal: (1) dificuldades e oportunidades; (2) intenções e *utterances*; (3) contexto; e (4) personalização.

Palavras-chave: Interação por voz; Alexa; Arquitetura; Problemas; Revisão sistemática da literatura.

Resumen: La interacción natural ha ganado espacio en los estudios de Interacción Humano-Computadora (HCI) porque es intuitiva, ya que se centra en las acciones espontáneas de los usuarios. La voz, por ejemplo, puede servir como un medio natural para interactuar con objetos mediáticos y, debido a esto, se han creado muchos modelos de interacción de voz en los últimos sesenta años. La diversidad y escala de publicaciones sobre interacción vocal hacen que el campo de investigación sea cada vez más complejo, haciendo necesaria una estructuración sistemática del campo. El objetivo de este estudio es analizar publicaciones científicas de los últimos seis años sobre problemas de interacción vocal, dando especial importancia al asistente virtual Amazon Alexa. El resultado del análisis sugiere 4 (cuatro) enfoques diferentes para los problemas de interacción vocal: (1) dificultades y oportunidades; (2) intenciones y declaraciones; (3) contexto; y (4) personalización.

Palabras clave: Interacción de voz; Alexa; Arquitectura; Problemas; Revisión sistemática de

[1] Post-graduate Program in Knowledge Engineering and Management; Federal University of Santa Catarina (UFSC); Florianópolis, Brazil; https://orcid.org/0000-0003-3955-0416; e-mail: luizadegraf@hotmail.com.
[2] Post-graduate Program in Knowledge Engineering and Management; Federal University of Santa Catarina (UFSC); Florianópolis, Brazil; https://orcid.org/0000-0001-6826-7180; e-mail: fapfialho@gmail.com.

la literatura.

1 INTRODUCTION

According to Dourish (1999), the expansion of tangible computing aims to eliminate coupling, emphasise directivity, and introduce new technologies that allow greater mobility and quality of HCI. Unlike old computational models, natural interaction opens horizons for prioritising the body, time, experience and the quality of human life. The emergence of voice user interfaces (VUIs) enabled spoken human interaction with computers and marked a major shift in the paradigm of Human-Computer Interaction. To discuss the subject of natural interaction and voice assistants, a brief introduction about the history of Human-Computer Interaction is primarily necessary. A brief summary of the evolution of computational models over the decades is presented below.

In the 1940s, the first computers required us to walk for metres and get into uncomfortable positions to execute commands. According to Dourish (1999), the performances were particularly electric and communication with humans usually took the form of flashing lights. Between the 1940s and 1970s, the creation of computers with stored programming, such as the IBM 360, transformed the electrical interaction into a symbolic one. According to Fisk (2005), the programs were constructed from punched cards and line-by-line prints. According to Bell (1968), around 1970, the sharing of computational resources among several users at the same time (time-sharing) was developed, as well as the possibility of using more than one program per computer, which made the HCI heavily based on human language, and scaled from symbolic to textual interaction. In 1980, we can observe the popularisation of the first graphical user interfaces, with screens capable of reproducing bitmaps, whose definition, according to Guibas and Stolfi (1982, p. 1, our translation), is “matrices of discrete values of intensity/colour”. The Xerox Alto computational model was a pioneer in using the desktop computer format, modernising textual interaction to graphical interaction. Notably, in 1973 (the same year as the launch of the Xerox Alto computer) the company Motorola launched the first cell phone, making communication mobility possible and profoundly changing the following decades. However, even with the advent of portable computers, in terms of body positioning, we are currently still prisoners of the desktop computer model (desktop) as we still spend a large part of our lives sitting in front of these machines. For Dourish (2001, p. 2), “the massive increase in computing power and the expansion in which we put that power to use - both suggest that we need new ways of interacting with computers, ways that are better suited to our needs and skills”.

This study seeks to understand which voice interaction problems are addressed in research on voice assistants, and aims to analyse scientific publications from the last decade on voice interaction problems, giving special importance to the problems of interaction with the voice assistant Amazon Alexa. The structure of the article begins with the theoretical framework on voice interaction, moving on to the presentation of methodological procedures, where the protocol used for the integrative literature review is exposed. Next, the results are shown, the results are discussed, and finally, comes the final considerations.

2 THEORETICAL FOUNDATIONS

In this article we present an analysis of voice interaction problems. Voice interaction creates machines that can recognize and respond to human speech through natural interaction, which in turn focuses on innate and instinctive human expressions towards some object to give users functional feedback. Natural interaction does not require the user to use external devices or learn procedures, as it has self-explanatory systems and a simple and intuitive interaction (Baraldi et al, 2009).

To categorise the selected publications, some understanding of the origin of voice interaction technology models was previously necessary. The following is a brief summary of the pioneers of voice interaction:

In 1939, the machine entitled The Voder (acronym for voice operation demonstrator) was exhibited at the World's Fair in New York. The Voder was the first synthesiser of the human voice assimilated as a musical instrument (Bush, 1945). In 1952, 13 (thirteen) years after the presentation of The Voder, three scientists from Bell Labs (K.H. Davis, Rulon Biddulph and Stephen Balashek) created the AUDREY machine (acronym for AUtomatic Digit REcognition), which is capable of recognizing spoken numbers (Pieraccini, 2012). The process of recognizing and reproducing human speech is a two-way street, so The Voder (synthesises) and AUDREY (recognises) machines are complementary to each other.

Currently, voice assistants are becoming widely used by the population, being incorporated into their homes and becoming part of their daily routines. Amazon Alexa, also known as Alexa, is a virtual assistant developed by the company Amazon, and is able to respond to voice commands and perform a multitude of tasks through its Skills (abilities), such as playing music, setting alarms, streaming podcasts, playing audio books, and delivering news on a variety of topics in real time, including politics, weather, and sports. The devicer also controls smart devices.

Recent advances in automatic speech recognition and natural language processing have enabled a new generation of robust speech interfaces (Lyons et al, 2016). However, such devices still do not fully exploit their potential and have serious problems that can reduce the naturalness of voice interaction. This article seeks to understand what are the main problems found in voice assistants.

3 METHODOLOGICAL PROCEDURES

This article was built through an integrative literature review on voice interaction problems. The integrative literature review is a review method that summarises the empirical or theoretical literature on the subject to provide a more comprehensive understanding of a given phenomenon (Whittemore, 2005). Once the theoretical foundations were described in the previous section and the research question determined, at the suggestion of the master's advisor, it was concluded that the base to be consulted would be the Association for Computer Machinery - ACM, due to containing the main and most cited articles in the sciences of Human-Computer Interaction.

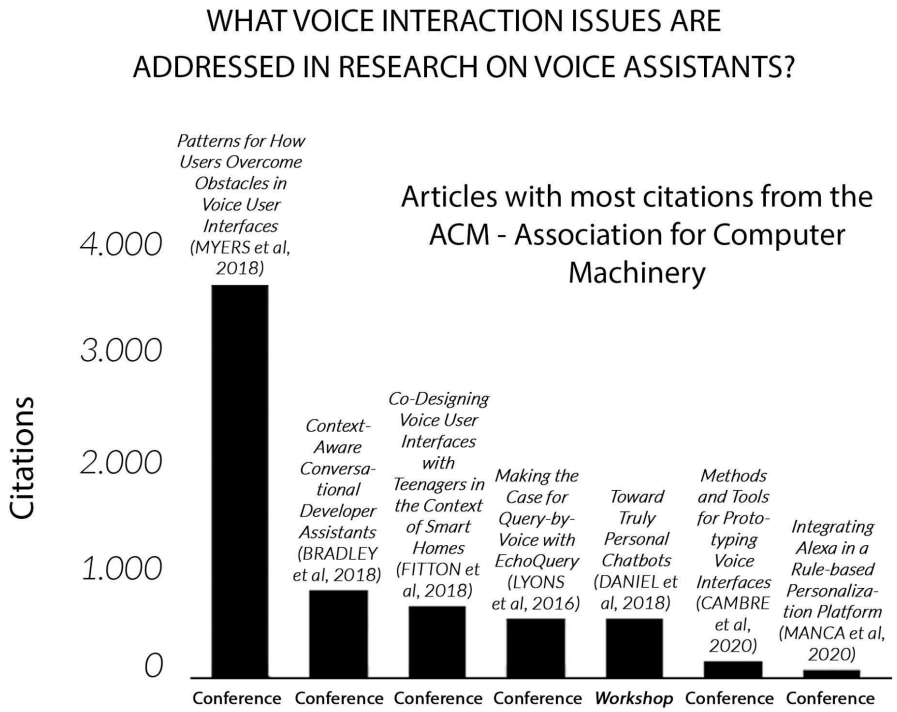
This article followed a protocol that guided the search and selection of articles. Articles from 2016 to 2022 were included, only peer-reviewed, only in English and only from conferences or workshops. Articles that did not have the terms “voice” were excluded; “voice interaction”; “Alexa”; “natural interaction”; “problem”; and “architecture”. The quality of selected publications was verified through collection in journals with a high impact factor. The data extraction strategy was carried out through a dense reading of the 7 (seven) selected articles and the outline of the most relevant parts for the review. The data analysis strategy was carried out by categorising the most interesting parts within 4 (four) different approaches to voice interaction problems. A table was composed with data extracted from each of the selected articles, containing the date of publication of the article; researchers' names; theoretical foundations of the research; research question, and place of publication of the article; the table that can be found in the next section.

4 RESULTS

The selection consisted of 28 (twenty-eight) different authors. The most cited articles in the ACM database, in descending order, were: “Patterns for How Users Overcome Obstacles in Voice User Interfaces”; “Context-Aware Conversational Developer Assistants”; “Co-Designing Voice User Interfaces with Teenagers in the

Context of Smart Homes"; "Making the Case for Query-by-Voice with EchoQuery"; "Toward Truly Personal Chatbots"; "Methods and Tools for Prototyping Voice Interfaces"; and "Integrating Alexa in a Rule-based Personalization Platform", as shown in the chart below:

Figure 1 - Research question and analysis of publication



Source: Author (2022).

The systematic search took place at the Association for Computer Machinery - ACM and selected articles from 2016 to 2022. The terms "voice" were used; "voice interaction"; "Alexa"; "natural interaction"; and "problem" which generated 568,258 articles as a result, and later added the term "architecture" which reduced the results to 289 articles.

After reading all the titles, the 20 (twenty) most interesting titles were selected, that is, those that contained words referring to the search. In the next step, the 20 (twenty) abstracts were analysed and separated in order of relevance to the research. Subsequently, 7 (seven) articles in which the word "Alexa" appears more frequently were selected for dense reading.

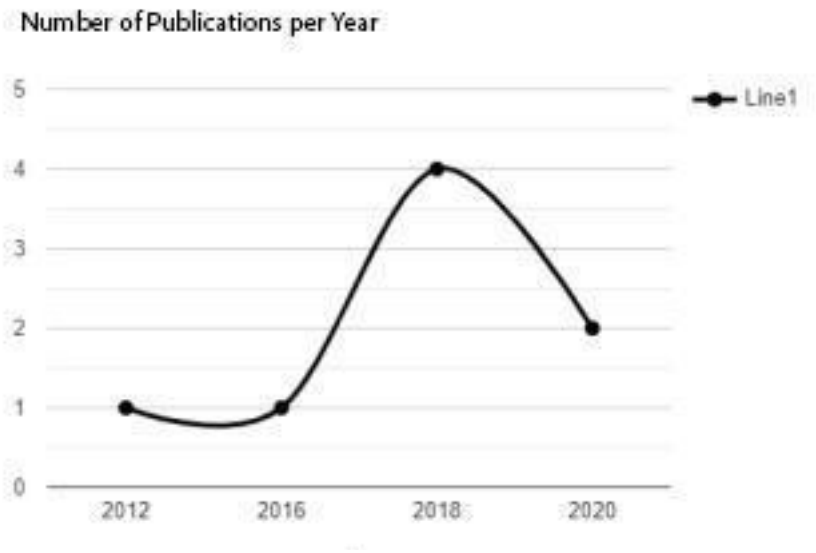
Due to its relevance to the topic, the book The Voice in the Machine by Roberto Pieraccini, Google Assistant engineering director, was incorporated as an extra reference. Thus, 1 (one) book and 7 (seven) articles were configured as a sample for this review. Of the total of 8 (eight) publications, 6 (six) are from conferences, 1 (one) is from a workshop, and 1 (one) is a book.

5 DISCUSSION

The search strategy used only the Boolean operator “and” and had as main components the keywords “voice”; “voice interaction”; “Alexa”; “natural interaction”; and “problem”, generating 568,258 articles as a result, later the term “architecture” was added, which reduced the results to 289 articles.

The book The Voice in the Machine by author Roberto Pieraccini was included as an extra reference due to its excellent structure and presentation on the subject. The book examines six decades of work on the scientific development of voice interaction technology. Most publications occurred in 2018, two publications are from 2020, one from 2016, and one from 2012, as shown in the following graph:

Figure 2 - Year of publications.



Source: Author (2022).

The analysis of the 8 selected works led to the organisation of the discussion into 4 topics. Articles were categorised according to their similarities and belonging themes. The result of the analysis was divided into 4 (four) distinct categories: difficulties and opportunities; intentions and utterances; context; and customization.

5.1 DIFFICULTIES AND OPPORTUNITIES

The fundamental question to be answered when we talk about the architecture of voice interaction is whether there are aspects of human speech that we still do not understand well enough to replicate in a machine (Pieraccini, 2012). Exposing large amounts of data to users

through voice assistants is a challenge, and providing an integrated view of different data through the use of speech is an even more complex task, which requires improving the presentation and analysis of this data by establishing principles for voice interface design (Daniel et al, 2018). Voice interfaces are enigmatic since the understanding and artificial reproduction of speech is a difficult task to be performed, especially due to paralinguistics (non-verbal aspects of language), such as intonation, tone of voice, speech rhythm, voice volume, among others (Cambre et al, 2020). Hyperarticulation of speech (speaking louder or slower) is often necessary to interact with the device, especially when there is noise in the environment or when music is playing (Myers et al, 2018). Suprasegmental features of speech can be extremely important to deliver the correct results to the user and avoid ambiguities, even for short sentences. Just one word can mean many different things depending on the paralinguistic aspects used (Pieraccini, 2012).

Despite the development difficulties encountered in recent decades, the exponential appearance of the ubiquity of voice interaction technology is remarkable (Cambre et al, 2020). Voice interfaces can bring many benefits to society, as they make room for users' peripheral cognition, and can make us more agile, smarter and better at a variety of tasks (Pieraccini, 2012). Voice interaction is an interesting alternative to interaction through the touch of graphical interfaces (Bradley et al, 2018) and when well designed, it can allow for more efficient multitasking (Fitton et al, 2018).

5.2 INTENTS AND UTTERANCES

Two terms are often used when talking about voice assistant architecture, intents and utterances. The intents are the intentions of the user's requests made to the voice interface, while the utterances are the words and utterances used to make requests to the interface (Myers et al, 2018). The interface interacts through the interpretation of statements, considering the history of past interactions, the user's current context, among other factors, and responding appropriately to requests (Pieraccini, 2012). The statements allow the interface to search for clarifications to optimise the fulfillment of the user's intentions, when there are incomplete statements or ambiguities in the request. (Cambre et al, 2020). Users of voice interfaces often simplify, change, or add more information to utterances to increase the success of the interaction. In addition to relying on graphical interfaces, users tend to restart, or completely give up on interactions when their requests are unsuccessful. Proper design of utterances can help reduce unsatisfactory interactions (Myers et al, 2018). The interface must

show interest in refining user requests and requesting clarifications on statements when necessary (Lyons et al, 2016).

5.3 CONTEXT

Everything said can be interpreted in different ways, and the correct interpretation is directly related to the context (Pieraccini, 2012). The user's context is directly related to the environment in which this user is found, and the mapping of this user's daily requests to the interface (Lyons et al, 2016). The context must be updated and memorised continuously in the background and to the same extent that the intents are digested by the interface, thus creating a relationship that allows the memorization of the user's usual environments and the goals of their desires in these different contexts, to for incremental improvement to take place (Bradley et al, 2018). Long-term memory plays a crucial role so that the assistant can distinguish users, support more than one, and provide personalised services (Daniel et al, 2018). Identifying the context opens possibilities for the interface to better understand the feelings of users and invite them to specific activities or control smart objects in the house according to the mood in which the user is in (Manca et al, 2020) .

5.4 CUSTOMIZATION

The voice interaction devices that are available to the population in the year 2022 are still not very customizable. Although voice assistants already cover a wide range of functions, they are still not personal enough and do not have intimate knowledge about users' preferences, needs and habits to be able to serve them with greater success (Daniel et al. , 2018). All voices are unique, like fingerprints (Pieraccini, 2012), but current assistants still do not support conversations with several different users by identifying each one's voice (Fitton et al, 2018). The creation of a deeper identity for the user would make it possible to design a customizable vocabulary incrementally structured, in addition to proactively offering appropriate suggestions to the user, as well as a guided improvement system (Lyons et al, 2016). Expanding the collection of user identification data can make it possible to improve the use of conversations and context data acquired over time, as well as this user's preferences. Through greater customization of voice interfaces, it is possible to consider developing a specific and customizable persona for each assistant, through a personality development extension. (Daniel et al, 2018).

6 FINAL CONSIDERATIONS

The paralinguistic aspects of speech and its ambiguities are still not well understood by voice interfaces. Users tend to simplify, change, or add more information to utterances to increase the success of the interaction. The usual contexts of requests are not identified by the interfaces and in certain environments there is a need for speech hyperarticulation. Exposing large amounts of data through voice is a complex challenge that requires profound improvements in voice interaction design. The collection of data on the specific profile of each user must be deepened so that the understanding of users' needs and feelings is improved. Voice interfaces do not support customizable services and vocabularies, nor do they identify different users, which makes it impossible to create specific and customizable personas for each assistant.

Voice assistants lack well-established design principles. Future research should make use of exploration design in order to establish design principles that consider the context and specific characteristics of each user and that make interactions more natural. Finally, it is recommended that voice interaction problems be seen as opportunities to improve the user experience.

The limitations of this article are several due to the context in which the research was developed and the complexity of the subject. This was the first research attempt made by a student at the beginning of the development of her master's thesis, and it presents structural problems. The findings of the review responded to the research question superficially, and it was necessary to complement the results through the book The Voice in the Machine by author Roberto Pieraccini. Criticisms of the method are related to the complexity of the topic and the search strategy, which must be rethought in order to find more qualified titles to answer the research question. The number of articles analysed should also be greater.

THANKS

This work was carried out with the support of the Coordination for Improvement of Higher Education Personnel - Brazil (CAPES) - Financing Code 001.

REFERENCES

Baraldi, S., Bimbo, A. D., Landucci, L., & Torpei, N. (2009). Natural Interaction. *Encyclopedia of Database Systems, 1880–1885.* https://doi.org/10.1007/978-0-387-39940-9_243

Bell, C. G., & Gold, M. M. (1972). An Introduction to the Structure of Time-Shared Computers. *Advances in Information Systems Science, 161–272.*

https://doi.org/10.1007/978-1-4615-9053-8_4

Bradley, N., Fritz, T., & Holmes, R. (2018). Context-Aware Conversational Developer Assistants. *2018 International Conference on Software Engineering* https://www.cs.ubc.ca/~rtholmes/papers/icse_2018_bradley.pdf

Bush, V. (1945, July). As We May Think. *The Atlantic.* https://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/

Cambre, J., & Kulkarni, C. (2020). Methods and Tools for Prototyping Voice Interfaces. *Proceedings of the 2nd Conference on Conversational User Interfaces.* https://doi.org/10.1145/3405755.3406148

Daniel, F., Matera, M., Zaccaria, V., & Dell'Orto, A. (2018). Toward truly personal chatbots. *Proceedings of the 1st International Workshop on Software Engineering for Cognitive Services.* https://doi.org/10.1145/3195555.3195563

Dourish, P. (1999). Embodied Interaction: Exploring the Foundations of a New Approach to HCI. *www.dourish.com.* https://www.dourish.com/embodied/embodied99.pdf

Dourish, P. (2001). Where the Action Is. *www.dourish.com.* https://www.dourish.com/embodied/

Fisk D. (2005). Programming with Punched Cards. *Columbia University.* http://www.columbia.edu/cu/computinghistory/fisk.pdf

Fitton, D., Read, J. C., Sim, G., & Cassidy, B. (2018). Co-designing voice user interfaces with teenagers in the context of smart homes. *Proceedings of the 17th ACM Conference on Interaction Design and Children.* https://doi.org/10.1145/3202185.3202744

Guibas, L., & Stolfi, J. (1982). A Language for Bitmap. *ACM Transactions on Graphics, Vol.1, No.3, July 1982, Pages 191-214.* Retrieved August 22, 2023, from https://www.cs.tufts.edu/~nr/cs257/archive/leo-guibas/language-bitmap.pdf

Ishii, H., Wisneski, C., Brave, S., Dahley, A., Gorbet, M., Ullmer, B., & Yarin, P. (1998). ambientROOM. *CHI 98 Conference Summary on Human Factors in Computing Systems.* https://doi.org/10.1145/286498.286652

Lyons, G., Tran, V., Binnig, C., Cetintemel, U., & Kraska, T. (2016). Making the Case for Query-by-Voice with EchoQuery. *Proceedings of the 2016 International Conference on Management of Data - SIGMOD '16.* https://doi.org/10.1145/2882903.2899394

Manca, M., Parvin, P., Paternò, F., & Santoro, C. (2020). Integrating Alexa in a Rule-based Personalization Platform. *Proceedings of the 6th EAI International Conference on Smart Objects and Technologies for Social Good.* https://doi.org/10.1145/3411170.3411228

Myers, C., Furqan, A., Nebolsky, J., Caro, K., & Zhu, J. (2018). Patterns for How Users Overcome Obstacles in Voice User Interfaces. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18.* https://doi.org/10.1145/3173574.3173580

The Voice in the Machine. (2012). *MIT Press.* Retrieved October 13, 2023, from https://mitpress.mit.edu/9780262533294/the-voice-in-the-machine/

Whittemore, R., & Knafl, K. (2005). The integrative review: updated methodology. *Journal of Advanced Nursing, 52(5), 546–553.* https://doi.org/10.1111/j.1365-2648.2005.03621.x