UX Evaluation Methodology for iTV: Assessing a Natural Language Interaction System

Jorge Abreu[1](✉), Juliana Camargo[1], Rita Santos[2], Pedro Almeida[1], Pedro Beça[1], and Telmo Silva[1]

[1] DigiMedia, Department of Communication and Art, University of Aveiro, 3810-193 Aveiro, Portugal
{jfa,julianacamargo,almeida,pedrobeca,tsilva}@ua.pt

[2] DigiMedia, Águeda School of Technology and Management, University of Aveiro, 3754-909 Aveiro, Portugal
rita.santos@ua.pt

Abstract. The user experience can be evaluated in different ways, from a combination of approaches and techniques capable of collecting insights about a particular product or system. In the TV ecosystem research field, verifying how users react to new features is an essential step to the integration of functionalities able to enhance the user experience. This is the case of systems based on interaction by natural language (NLI), which have the potential to allow for more simplified navigation (based on conversational dynamics). However, although the spoken interactions are relevant to optimize the consumption of television content, it is essential to identify how they are received and understanded by the user. In this context, this empirical study sought to analyze the experience of using an NLI system controlled by a mobile application. The evaluation was performed employing an open methodology that considers instrumental and non-instrumental qualities of the application, as well as emotional dimensions. This approach was specifically developed for UX analysis of systems and applications related to the TV ecosystem, having recorded positive results in previous studies. In the study here presented, the methodology revealed again to be suitable as it was possible to identify failures and opportunities to improve the assessed NLI system and, finally, to verify that the voice interaction system allowed users a more optimized and accessible experience.

Keywords: User Experience · Evaluation · Triangulation · Methodology · iTV · NLI system · Voice interaction

1 Introduction

Evaluating the user experience (UX) is an increasingly relevant task, often reflected in academic studies, [1], being one of the main reasons for that the set of results that a well-structured analysis can offer.

This evaluation goes beyond the observation of the responses of individuals to the anticipated use of a product, system or service [2]. With a suited methodology, it is

M. J. Abásolo et al. (Eds.): jAUTI 2020, CCIS 1433, pp. 149–161, 2021.
https://doi.org/10.1007/978-3-030-81996-5_11

possible to detect emotional and hedonic characteristics, such as aesthetics, stimulation and identification, which are extremely relevant in the interactions between humans and systems [3].

Therefore, an adequate UX evaluation should not be limited to the usability dimension, that is, to the evaluation of operational tasks when using a product or service. For being a global experience, it also needs to be evaluated globally, using a methodology capable of identifying all aspects involved in the user's journey. Examining this path in detail is a multidisciplinary activity, encompassing cognitive sciences, psychology, engineering and design [2].

In the interactive television (iTV) field, the central theme of this study, UX evaluations are extremely important to validate new features and improve technologies. This is the case, for example, of natural language interaction (NLI) systems to operate an iTV solution. Although it is possible to use voice assistants to perform actions on TV, the true potential of spoken interactions is not yet widely explored in this context. That is, instead of enabling conversational dynamics (the main purpose of this type of system), interactions have been limited to using voice commands to swap channels or increase volume, for example. This scenario makes UX assessments important tools for making technical advances capable of turning the interaction more user-friendly and anthropomorphic. These are two characteristics considered important for a satisfactory UX [4].

In this sense, the present study is dedicated to describing an open UX evaluation methodology for the iTV domain, how it was applied to assess an NLI system specifically designed for an iTV commercial platform, and the correspondent results. This evaluation approach, already used in previous studies [5–7] performed by the Social iTV research group (http://socialitv.web.ua.pt), is based on a triangulation of free questionnaires combined with a semi-structured interview, being adaptable to a range of applications belonging to the field of the TV ecosystem. The methodology was specified by the authors to evaluate the perspectives of users on the instrumental and non-instrumental qualities of the application, as well as the emotional reactions aroused by the episodic/cumulative UX [8].

To present the context and gathered indicators, this paper is structured as follows: Sect. 2 presents a set of instruments used to evaluate UX systems focused on human-computer interaction, unfolding to describe the approach used in the present study; the prototype of the NLI system is presented in Sect. 3; Sect. 4 details the procedures, sample, and results obtained; discussion of the results is held in Sect. 5; and finally, the conclusions are presented in Sect. 6.

2 UX Evaluation methodologies for iTV Applications

2.1 Overview

Although there is no single definition about UX, the ISO9241-110: 2010 (clause 2.15) classifies the concept as "a person's perceptions and responses as a result of the early use and/or use of a product, system, or service" [9].

UX evaluation tends to be complex as it examines different aspects arising from the use of a product or software [10] and, consequently, there is no single formula applicable to all scopes.

The dimensions to be evaluated vary according to each case or test area, as they must be appropriate and relevant to the contexts in which they are applied. In order to evaluate the UX, evaluators have at their disposal a range of methods, approaches and scales, tested and recognized over the years [2]. As the UX is by its nature a complex matter, involving various aspects and requiring different types of responses [2], its evaluation is not limited to the usability and performance of interactive solutions, covering also non-instrumental qualities such as aesthetics, stimulation, and identification; emotional reactions, such as, pleasure, attraction and control; and timeless practice [17, 19].

An analysis of UX evaluations studies related with iTV applications allowed to identify that the following free established methods have been regularly used:

1. Self-Assessment Manikin (SAM): a non-verbal pictorial assessment method that assesses levels of satisfaction, motivation and control.
2. AttrakDiff: a questionnaire based on a semantic differential scale that evaluates two components of an interactive application or product – Pragmatic and Hedonic Quality.
3. SUXES: an evaluation method for collecting subjective metrics with user experiments. It captures the expectations and experiences of individuals, making it possible to analyze the state of the application and its methods of interaction.

As the usability of the iTV applications may have an impact on the perceived non-instrumental dimensions of the UX, it is also frequent to resort to the combined use of the System Usability Scale (SUS) - a set of ten simple questions, answered by means of a five-point Likert scale, related to the overall usability of the application.

The following studies sought to evaluate the UX of products and solutions that enable human-computer interaction, most of which belonging to the context of iTV. Almost all the articles mentioned here have used more than one instrument to evaluate the user experience.

For example, Lee et al. [11] used the SAM scale combined with a semantic differential questionnaire to analyze the felling and the perceived quality of the interactive features embedded in the television.

Similar procedures were performed by Ludwig et al. [12], Rodrigues et al. [13] and Pailleur et al. [14] identifying emotional aspects related to intelligent systems from the combination of more than one instrument, such as the SAM questionnaire and semi-structured interviews.

An example of applying the AttrakDiff scale can be seen in [15], which sought to identify, among the elderly, pragmatic, hedonic and attractive components in the use of three different TV remote controls: a prototype designed especially for the elderly, the Tekpal model, aimed at senior citizens, and a traditional model, commonly used in a daily basis. The evaluation made it possible to identify the most appropriate remote control for the target users. Contrary to most of the other studies, here only the AttrakDiff was used to evaluate the user experience.

The SUXES evaluation method, in turn, was used by Turunen [16] to evaluate the UX of various modes of interaction of a home entertainment system controlled by a mobile phone. Such a method was able to collect expectations and experiences, making it possible to analyze the state of the application and its interaction methods (and compare results). The researchers combined the SUXES method with questionnaires applied before and after the test, where they requested an overall assessment of the user experience [16].

Some researchers, such as Ouyang et al. [18], used SUS as one of the methods of the UX evaluation described in the study. In this case, the authors evaluated UX in three stages. In a first moment: the participants were asked to complete a basic questionnaire about their background and daily TV use. Second, a think-aloud demonstration was presented to participants before they attempted to complete the assigned tasks. Third, participants were asked to complete the SUS questionnaire [18].

Given some of the gaps found, such as lack of clarity regarding the dimensions that were evaluated by the researchers and the use of a dominant method, a more comprehensive methodology was proposed by the Social iTV research group of the University of Aveiro (described in the following section).

2.2 Evaluation Methodology Proposal

The methodology used in this study to evaluate the experience associated with the natural language interaction with the TV set draw on the dimensions identified on a literature review oriented to the TV ecosystem, conducted by Bernhaupt and Pirker [3]. These dimensions are: i) stimulation (describes the extent to which a product can meet the user's needs with attractive functions, interactions, and content); ii) identification (a product's ability to allow a user to identify with it); iii) emotional (feelings and emotions triggered by the experience, with emphasis on satisfaction, motivation, and control); and iv) visual/aesthetic (levels of experience attractiveness). The proposal follows the CUE (Components of User Experience) model [20], which reinforces the importance of considering emotions and perceptions of instrumental qualities in articulation with the ones resulting from non-instrumental qualities.

This structure was the starting point for the development, by Abreu, Almeida, and Silva [5], of an open methodology based on the triangulation of UX instruments capable of evaluating precisely the dimensions highlighted by Bernhaupt and Pirker [3]. With this free and open methodology, the non-instrumental dimensions of stimulation and identification are evaluated using the Hedonic Quality components of the AttrakDiff questionnaire (HQ-S and HQ-I); the emotional reactions (satisfaction, motivation, and control) rely on the SAM questionnaire, whereas the visual/aesthetic dimension is obtained by the attractiveness value of the AttrakDiff (ATT). To weigh the instrumental qualities (which can give relevant insights on how the perception of usability of the application relates to its UX), the methodology proposed by the team is supported on the SUS scale and on the pragmatic dimension of the Attrakdiff questionnaire. Finally, but no less important, to complement the data obtained a semi-structured interview is carried with each of the participants who tested the system (Fig. 1).

The evaluation process is divided into three stages, the first being the preparation (which comprises the definition of objectives, the preparation of the setup, the definition of variables and the preparation of the instruments). Then, after the episodic/cumulative experience, the application of the data collection tools take place (in the order of SAM, SUS, Attrakdiff questionnaires complemented by the interview). Finally, the data collected is analyzed so that the conclusions can be addressed.

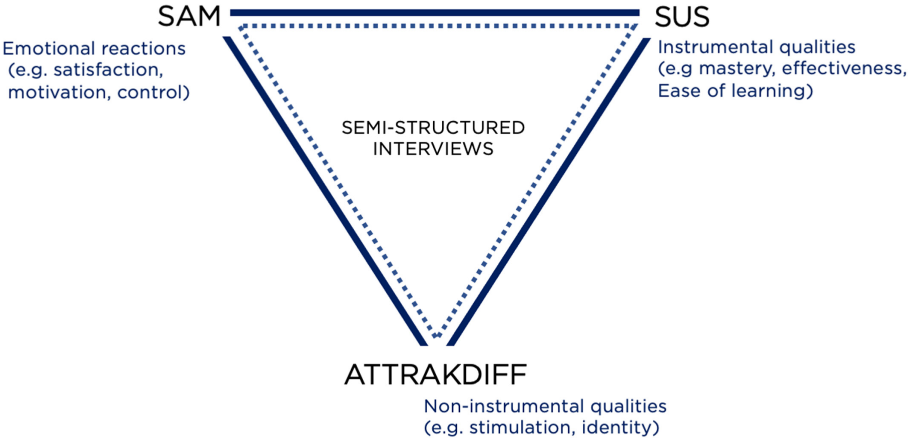


Fig. 1. Methodology based on the triangulation of UX scales.

This methodology was used in previous studies, having contributed significantly to evaluate experiences resulting from the use of: i) a mobile application centered on the recognition of audio and video of a set of interactive systems related to real-time TV services [5]; ii) an application (second screen type) aimed at the search for television content [6]; and iii) an advanced TV content aggregation User Interface that allows to offer TV and OTT contents at the same level [7]. These studies obtained consistent results, allowing to know the overall UX in relation to the evaluated systems and reinforcing the validity of the used methodology.

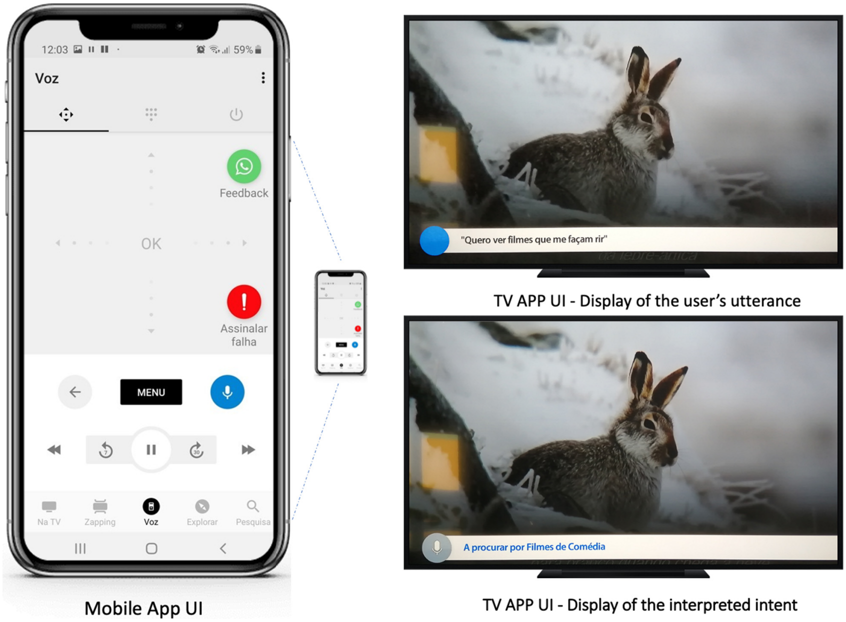
In the current evaluation, the authors sought to evaluate the referred dimensions related to the UX of a NLI system used in the TV context, with the aim of identify failures, correct, and add significant improvements.

3 Prototype of a Natural Language Interaction System for iTV

A voice interaction system, specifically developed for the Portuguese television context, was integrated into the main IPTV service provider in Portugal. Due to technical reasons, it was decided to upgrade an existing mobile application available to the company's customers, that worked as a virtual remote control, with a specific area integrating the NLI feature.

The interaction process starts when the user presses and hold a button (identified with a microphone icon – see Fig. 2) to utter the desired action using natural language. The captured audio is converted to text by a cloud-based Automatic Speech Recognition system (ASR), which processes and sends the spoken phrase to the iTV Set-Top Box (STB). The interface was designed to offer a fluid, natural and clear experience, including strategic resources, such as icons and phrases capable of guiding the user [16].

The user speaks and its utterance is immediately displayed in the iTV User Interface (UI), followed by a message on the TV screen indicating the correspondent intent (e.g., Searching for comedy movies). The results are presented below the message accompanied by the corresponding thumbnails. If there is a misunderstanding of



"Quero ver filmes que me façam rir"

A procurar por Filmes de Comédia

12:03
59%
Voz
Feedback
OK
Assinalar falha
MENU
Na TV
Zapping
Voz
Explorar
Pesquisa
Mobile App UI
TV APP UI - Display of the user's utterance
TV APP UI - Display of the interpreted intent

Fig. 2. App UI (left) and TV app UI when searching for comedy movies (right).

the utterance by the NLU, two possible actions may occur: either the system performs an action that is not as expected or issues a decoy, such as "Sorry, I still can't help you with what you asked for".

In addition, to make the UX more contextualized [7], the user is also able to immediately report eventual errors through the mobile app, using the "flag failure" (red) button, or interact through the "feedback" (green) button, that starts a conversation on WhatsApp (with a member of the research team) enabling the user to address issues raised by its momentary UX [5].

4 UX Evaluation Process

The UX evaluation of the NLI system was carried out in a real context of use, in a Field Trial (FT), building on the potential advantages of revealing problems that would not appear in a laboratory and providing a more realistic perspective of the commonly used phrases [19]. Its main objective was to verify the viability of the solution and to analyze improvements to be made to enhance its UX.

As proposed by the team's previous work [5], the scales used (SUS [21] and AttrakDiff) were a version translated to Portuguese and made available, along with the SAM scale, in a single (online) questionnaire for the participants.

After that, the data collected by these instruments used in the evaluation was complemented by a semi-structured interview, which had as objectives: i) to collect the opinion of the participants regarding the functionalities of the NLI system; ii) to identify the possible actions to be taken to improve the overall solution, and iii) to understand the level of willingness of participants in using the solution.

4.1 Procedures

The field tests were performed by 20 users between October 2019 and April 2020, spanning in a total of 169 days. Participants used the application in their homes for daily TV consumption activities. In addition, they were encouraged to test specific functionalities through challenges (with pre-defined themes, such as asking to see comedy films, finding content of actors and actresses, increasing or lowering the volume and finding programs using similar names, among others) sent on a weekly basis by e-mail. Such autonomy given to users made it possible to assess, in a more reliable way, the experience of using the proposed solution.

After the system testing period, the UX evaluation was carried out, being this a fundamental procedure for the validation of the prototype. Users were asked to answer the online questionnaire - built from the triangulation of scales (SAM, SUS, and AttrakDiff). The questionnaire sent to the participants had the aim of identifying the following global aspects: emotions triggered using the application, usability of the natural language interaction system, specific opinions about the natural language interaction system and suggested improvements.

Then, among the 20 selected participants, the 11 most active participants were invited to participate in a semi-structured videocall interview to identify and clarify aspects relevant to the study. Some examples of questions that were asked to the participants were: “would you use the system on a daily basis?”; “what were the main problems encountered?”; “what do you believe can be improved in the application?”; and “would you use the app in place of the remote control?”. This step was decisive to the evaluators, enable them to gather more consolidated opinions about the topics considered relevant for the UX evaluation.

4.2 Sample Characterization

A non-probability, by convenience, sampling was used and the prior knowledge of iTV apps was considered an inclusion criterion. Among the 20 selected participants, 75% (15) were men and 25% (5) women, with an average age of 44 years. Regarding the level of education, 45% (9) have a degree, 50% (10) a master’s degree and 5% (1) a doctorate.

Among the devices, regular TVs (connected to STB) are used on a daily basis by 90% (18) of the participants, followed by Smart TVs (5–25%), applications to control the TV (5–25%), Media Players (2 -10%) and vi assistant (2–10%).

The average consumption of television was 1 h and 37 min a day. Regarding the daily frequency of use of the TV features, 50% (10) use automatic recordings, 40% (8) pause television content, 15% (3) resort to recording content, 10% (2) to TV-guide and 5% (1) to TV content search. Regarding the use of voice interaction devices, only 25% (5) stated that they had already tested or had done it daily. The assistants that appeared in the responses were Google Home, Android Auto, Google Assistant and Smart TVs with integrated voice interaction.

Then, an analysis of the volume of interactions generated by the 20 individuals belonging to the sample was performed. From this group, we chose the 11 most active for semi-structured interviews, considering the quality and frequency of interaction performed during the period that covered the tests.

4.3 Results

The SUS scale identified data on instrumental qualities, namely on efficiency, effectiveness and ease of learning of the NLI system. The prototype obtained a score of 82 (on a scale from 0 to 100), which indicates that the average value of the participants' subjective perceptions about their usability is considered "Good" (Fig. 3.) The fact that the prototype underwent a process of continuous improvement throughout the tests contributed to this positive score.

Faced with a 5-level Likert scale (SUS scale), in which 1 – "I totally disagree" and 5 – "I totally agree", the results showed that participants would like to use the NLI system frequently (a = 4.3), found it easy to use (a = 4.1), consider that its functionalities were well integrated (a = 3.9) and quickly learned how to use it (a = 4.3). These indicators corroborate the opinions collected in the semi-structured interviews. Among the 11 respondents, eight said they would use the system daily and three would adopt it for activities that require greater cognitive load, such as searching for specific content.

Regarding privacy and security, the average stood at 4.1, suggesting that participants did not feel significant concerns about these themes. In the interviews, five persons stated that the use of the application can raise privacy issues for the target public, such as undue access to users' data, although this is not a problem that directly affects them. Five participants said that the NLI-system does not generate controversies about privacy. Only one showed insecure in relation to this theme, saying that "*I have a concern for systems that are always listening to us. There must be clarity in the policy for accessing customer data*".

The less positive score was recorded in the phrase "I think that this product had many inconsistencies", a clear reflection of some failures that were found during the use of the system (for example, the request "*turnoff the (set-top) box*" was not working, and the request "*forward*" was changing to the next channel instead of forwarding the content).

When comparing the results of participants with previous experience (n = 5) and without previous experience (n = 15) in using voice interaction systems/devices, it was verified that the prototype score was slightly higher among participants with previous experience (85 out of 100). However, the prototype obtained a score of 81 by the participants without such experience, which reveals that prototype commands are probably intuitive.

Regarding the SAM scale, it was found that, on average, the parameters "satisfaction" (3.85), "motivation" (3.65) and "control" (3.5) presented a positive score (in a scale of 1 to 5) but there is still room for improvement in terms of user-system interaction. In the semi-structured interviews, the 11 participants reported that they had some kind of problem during the experiment, and the major complaints were related to ambiguous commands and the system's difficulties in perceiving commands in English (both failures/difficulties mentioned four times). Another feature that troubled users was the need to hold the phone button to activate all voice commands (problem mentioned by three participants). Two

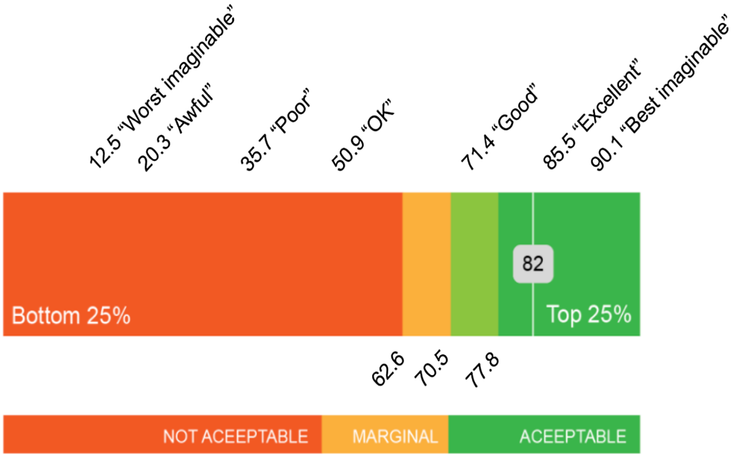


Fig. 3. Overall SUS Scale result.

of them claimed to be uncomfortable holding the button while saying the entire sentence and one reported that he had difficulties noticing when it was the exact time to release it. Although they were mentioned, these problems were not considered serious by the users interviewed, which was also reflected in the results of the questionnaires.

The "motivation" factor confirmed some desire to use the system. However, such a willingness to use the system may not be sufficient to replace the traditional remote control. The results of the semi-structured interviews showed that of the 11 participants, two stated that they would adopt the application only in conjunction with the remote control, because they consider that some tasks are easier to perform with it. One of the participants stated that it would adopt only if the app had a more practical access, because he felt uncomfortable with having to use it via the smartphone.

In view of an analysis and comparison of the data collected from the participants with (n = 5) and without previous experience (n = 15) of use voice interaction systems/devices, the results showed that in the parameters of satisfaction and motivation the scores were higher in the participants without previous experience, which may indicate that the "novelty factor" had important implications in these collected results.

Regarding the control parameter, the results showed that the participants with previous experience gave a higher score (4.00), compared to the participants without previous experience (3,33). This indicates that having a general and prior understanding of the capabilities and functionalities of voice assistants can improve and ease the experience of interaction between user-NLI system.

Regarding the AttrakDiff scale (−3 to 3), the pragmatic quality presented the lowest classification (1.05), and the aesthetic attractiveness presented the best classification (1.91). The hedonic qualities of identification and stimulation presented classifications of 1.24 and 1.56, respectively.

As for the pairs of adjectives, the pleasant-negligible and low cost-premium pairs presented negative classifications, with impact on both the pragmatic dimension and the identification dimension, which may indicate critical characteristics that should be improved. According to the results of the qualities, hedonic and pragmatic, the prototype was positioned between the "Self-oriented" and "Desired" quadrants (Fig. 4).

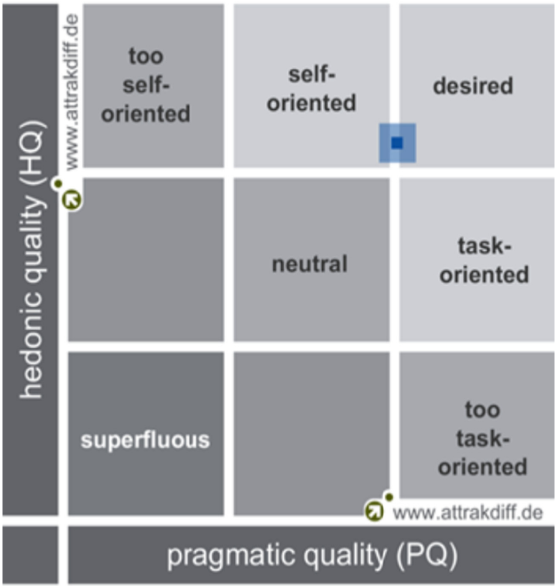


Fig. 4. AttrakDiff scale results.

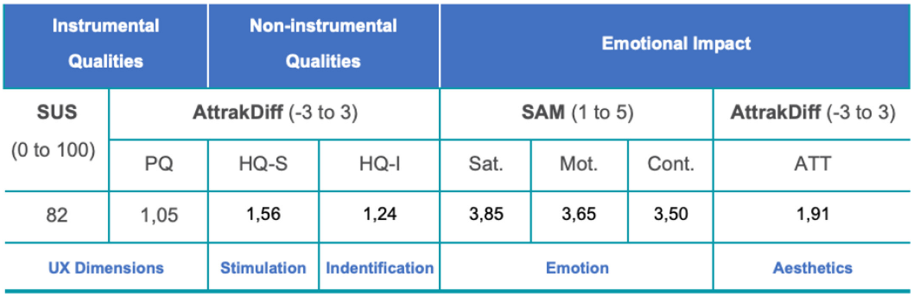
In the comparison between participants with previous experience (n = 5) and those without experience (n = 15), it was found that participants with previous experience attributed higher classifications in the parameters of stimulation (1.91) and attractiveness (2,20), while the participants without previous experience attributed a higher score in pragmatic quality (1,18).

Regarding the hedonic quality of identification, both gave the same score (1,23). Within these parameters, the attractiveness obtained better scores in both groups.

The global result of the AttrakDiff scale is in accordance with the opinions of the semi-structured interviews, in which the participants stated that they would like to use the system, mainly because it simplifies the experience of television consumption.

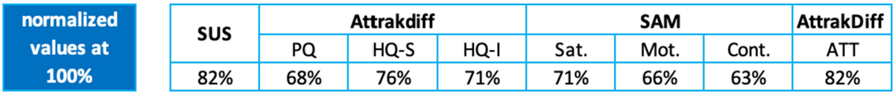
5 Discussion

In the triangulation of the three scales (Figs. 5 and 6), it is possible to realize that the prototype is overall satisfactory in terms of UX.



Instrumental Qualities		Non-instrumental Qualities		Emotional Impact			
SUS (0 to 100)	AttrakDiff (-3 to 3)			SAM (1 to 5)			AttrakDiff (-3 to 3)
	PQ	HQ-S	HQ-I	Sat.	Mot.	Cont.	ATT
82	1,05	1,56	1,24	3,85	3,65	3,50	1,91
UX Dimensions		Stimulation	Indentification	Emotion			Aesthetics

Fig. 5. Global scores of field tests - triangulation of SUS, SAM and AttrakDiff.



normalized values at 100%	SUS	Attrakdiff			SAM			AttrakDiff
		PQ	HQ-S	HQ-I	Sat.	Mot.	Cont.	ATT
	82%	68%	76%	71%	71%	66%	63%	82%

Fig. 6. Normalized scores of field tests.

In relation to the instrumental qualities of the prototype, the scores on the Pragmatic Quality of the AttrakDiff scale and that of the SUS scale reflect the user's comfort in relation to the use of the product (according to the evaluation scale, the average usability value considered good is 72.40 points). However, although the obtained value (82) is a good score, it indicates that there is still room for improvement, corroborating the opinions collected in the semi-structured interviews. Individual conversations with participants allowed us to detect a set of 11 improvements, such as "finer searches", "zapping (channel surfing) like that performed by the remote control", "direct access to content already seen" and "implementation of the view command from the beginning".

The less positive score in SUS was recorded in the phrase "I think that this product had many inconsistencies", which may have influenced the "control" dimension of the SAM scale, which obtained the lowest mean (3.50) compared to the other two emotional reactions: satisfaction and motivation. On the other hand, the fact that both obtained more positive scores indicates that there was a positive affective relationship regarding the use of the prototype, which was also identified with the result of the attractiveness value of the Attrakdiff scale (ATT).

Summing up, the positive feeling regarding usability may have contributed to increase the levels of satisfaction, motivation and stimulation. On the other hand, the flaws found throughout the FT probably interfered with important aspects such as control and simplicity.

Based on these results, it can be verified, therefore, that the applied methodology allowed to evaluate all aspects considered important for the television ecosystem and helped to improve the system for the next steps, making it more robust.

6 Conclusion

The iTV ecosystem is leaning towards the increasing use of voice interaction features [22], since spoken interactions have the potential to ensure a more user-friendly and practical UX [23]. Therefore, based on this scenario, we sought to perform a thorough evaluation of the experience of using a prototype that allows natural voice interaction, from a mobile application. The main objective was to make effective contributions to the TV ecosystem, testing, once again, an approach capable of identifying relevant failures and improvements.

Using an open methodology centered on the triangulation of UX scales and complemented with interviews, it was possible to evaluate the perspectives of users on the instrumental and non-instrumental qualities of the prototype, as well as the emotional reactions triggered by its UX.

The quantitative data supported the user satisfaction, especially in relation to fundamental aspects/dimensions for the television context, such as innovation, aesthetics, comfort and intention of use (following this experimental phase). Semi-structured interviews allowed to qualitatively assess the main obstacles and positive aspects of the experience of using the proposed solution.

From the methodology adopted it was possible to identify real problems arising from the experience of using the system. In addition, the adopted methodology provided essential insights to support the idea that natural voice interaction can be well accepted by TV users.

After the evaluation cycle it was possible to move towards a more stable and complete version of the presented solution. And, in this sense, the fact that the results were obtained from the actual use of the developed prototype contributed to reaffirm the relevance of the methodology to future projects carried out within the TV ecosystem.

Finally, it is also important to highlight some limitations of the study that may have interfered with the results obtained, such as sample size, gender disparity and lack of people without high academic training.

References

1. Lallemand, C., Koenig, V.: Measuring the contextual dimension of user experience: development of the user experience context scale (UXCS). In: ACM International Conference Proceeding Series (2020)
2. Pettersson, I., Lachner, F., Frison, A., Riener, A., Butz, A.: A bermuda triangle? - A review of method application and triangulation in user experience evaluation. In: Conference on Human Factors in Computing Systems – Proceedings (2018)
3. Bernhaupt, R., Pirker, M.: Evaluating user experience for interactive television: towards the development of a domain-specific user experience questionnaire. In: Kotzé, P., Marsden, G., Lindgaard, G., Wesson, J., Winckler, M. (eds.) INTERACT 2013. LNCS, vol. 8118, pp. 642–659. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40480-1_45
4. Bahlenberg, R., Yan, X.: Anthropomorphic design and anticipated user experience. In: Frontiers in Psychology (2019)
5. Abreu, J., Almeida, P., Silva, T.: A UX evaluation approach for second-screen applications. In: Abásolo, M.J., Perales, F.J., Bibiloni, A. (eds.) jAUTI/CTVDI -2015. CCIS, vol. 605, pp. 105–120. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-38907-3_9
6. Ferraz de Abreu, J., Almeida, P., Beça, P.: InApp questions – An approach for contextual evaluation of applications. In: Abásolo, M.J., Almeida, P., Pina Amargós, J. (eds.) jAUTI 2016. CCIS, vol. 689, pp. 163–175. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-63321-3_12

7. Velhinho, A., Fernandes, S., Abreu, J., Almeida, P., Silva, T.: Field trial of a new iTV approach: the potential of its UX among younger audiences. In: Abásolo, M.J., Silva, T., González, N.D. (eds.) jAUTI 2018. CCIS, vol. 1004, pp. 131–147. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-23862-9_10
8. Roto, V., Law, E., Vermeeren, A., Hoonhout, J.: User experience white paper - Bringing clarity to the concept of user experience. In: Outcome of the Dagstuhl Seminar on Demarcating User Experience, Germany. Seminar (2011)
9. ISO 9241-210.: Ergonomics of Human-System Interaction – Part 210: Human-centered Design for Interactive Systems (formerly known as 13407). International Standardization Organization (ISO), Switzerland (2010)
10. Law, E.: The measurability and predictability of user experience. In: Proceedings of the 2011 SIGCHI Symposium on Engineering Interactive Computing Systems. EICS 2011 (2011)
11. Lee, S., Yun, M.: Interactive TV user experience in behavioral situations. In: Multi Conference on Computer Science and Information Systems, MCCSIS 2019 - Proceedings of the International Conferences on Interfaces and Human Computer Interaction 2019, Game and Entertainment Technologies 2019 and Computer Graphics, Visualization, Comp (2019)
12. Ludwig, R., Bachmann, A., Buchholz, S., Ganser, K., Glänzer, D., Matarage, A.: How to measure UX and usability in today's connected vehicles. In: Ahram, T., Taiar, R., Langlois, K., Choplin, A. (eds.) IHIET 2020. AISC, vol. 1253, pp. 17–21. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-55307-4_3
13. Rodrigues, A., Machado, B., Almeida, M., Abreu, J., Tavares, T.: Evaluation methodologies of assistive technology interaction devices: a participatory mapping in Portugal based on community-based research. In: IHC 2019 - Proceedings of the 18th Brazilian Symposium on Human Factors in Computing Systems (2019)
14. Le Pailleur, F., Huang, B., Léger, P.-M., Sénécal, S.: A new approach to measure user experience with voice-controlled intelligent assistants: a pilot study. In: Kurosu, M. (ed.) HCII 2020. LNCS, vol. 12182, pp. 197–208. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-49062-1_13
15. Mehrotra, S.: Potmote: a TV remote control for older adults. In: ASSETS 2018 - Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility (2018)
16. Turunen, M.: User expectations and user experience with different modalities in a mobile phone controlled home entertainment system. In: ACM International Conference Proceeding Series (2009)
17. Guerino, G., Valentim, N.: Usability and user experience evaluation of natural user interfaces: a systematic mapping study. In: IET Software (2020)
18. Ouyang, X., Zhou, J.: How to help older adults move the focus on a smart TV? Exploring the effects of arrow hints and element size consistency. Int. J. Hum. Comput. Interact. **35**, 1420 (2019)
19. Hassenzahl, M.: The Thing and I:Understanding the Relationship between User and Product in Funology: From Usability to Enjoyment (2003)
20. Thüring, M., Mahlke, S.: Usability, aesthetics and emotions in human-technology interaction. Int. J. Psychol. **42**, 253 (2007)
21. Martins, A., Rosa, A., Queirós, A., Silva, A., Rocha, N.: European Portuguese validation of the system usability scale (SUS). Procedia Comput. Sci. **67**, 293 (2015)
22. Silva, T., Almeida, P., Abreu, J., Oliveira, E.: Interaction paradigms on iTV: a survey towards the future of television. In: 9th International Multi-Conference on Complexity, Informatics and Cybernetics. IMCIC, pp. 18–23 (2018)
23. Kocaballi, A., Laranjo, L., Coiera, E.: Understanding and measuring user experience in conversational interfaces. Interact. Comput. **31**, 192–207 (2019)