Assessing a Technology for Usability and User Experience Evaluation of Conversational Systems: An Exploratory Study

Guilherme Corredato Guerino[1], Williamson Alison Freitas Silva[2], Thiago Adriano Coleti[2] and Natasha Malveira Costa Valentim[1]
[1]*Federal University of Paraná, Curitiba, Brazil*
[2]*State University of Paraná, Apucarana, Brazil*

Keywords: Usability, User Experience, Conversational Systems, Evaluation, Voice-based Interaction.

Abstract: Conversational Systems (CS) are increasingly present in people's daily lives. CS must provide a good experience and meet the needs of its users. Therefore, the Usability and User Experience (UX) evaluation is an appropriate step before making CSs available to society. To guide developers to identify problems, improvement suggestions, and user perceptions during CSs development, we developed a technology named Usability and User Experience Evaluation of Conversational Systems (U2XECS). U2XECS is a questionnaire-based technology that provides Usability and UX statement specifics to evaluate CSs. We conducted an exploratory study performed to evaluate and evolve U2XECS. Our results evidenced positive points of U2XECS related to ease of use, usefulness, and intentions to use. Moreover, we identified opportunities for improvement in U2XECS, such as ambiguous statements that generated misinterpretations in subjects.

1 INTRODUCTION

Conversational Systems (CS) are increasingly present in our daily lives and because of this they receive investments from the industry. It is estimated that only the Intelligent Virtual Assistants (IVAs) market will reach $45.1 billion in 2027 (Grand View Research, Inc, 2020). The high interest in this type of system is because it is widely used in applications for smartphones (e.g., Alexa), physical devices for home automation (e.g., Google Home), integration with car systems (Google Assistant integration), and even smart glasses (Amazon Echo Frame).

CS also are covered in Human-Computer Interaction (HCI) research. The HCI area has provided several contributions in the CS topics (Porcheron et al., 2018) in several contexts and users profile, such as elderly users (Trajkova and Martin-Hammond, 2020), and children (Lovato and Piper, 2015), as well as case studies (Candello et al., 2019) and literature reviews (Kocaballi et al., 2020).

In the context described above, CS should be useful, easy to use, and enjoyable to be accepted in society. Then, it is essential to evaluate the quality of CSs. In literature, one way to evaluate the quality of systems is through Usability and User Experience (UX) evaluation. Usability and UX evaluation allows developers to have a perception - positive or negative - of the developed system before making it available to market. From this step and the results obtained, developers can improve negative points and explore better positive topics.

Kocaballi et al. (2018) reveal that there are several interpretations of Usability and UX for the CS since UX is used: (i) as Usability; (ii) as something beyond Usability; or (iii) as user satisfaction. In this paper, we interpreted Usability according to ISO 9241-210 (2019), being "the extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use." UX is related to "user's perceptions and responses that result from the use and/or anticipated use of a system, product or service" (ISO 9241-210, 2019). While the Usability evaluation verifies behavioral goals (such as efficiency and effectiveness), the UX evaluation understands the concepts related to users' subjective perceptions (such as motivation and emotion). Therefore, technologies[1] that evaluate Usability and UX could provide to the developer a vision that encompasses systems' behavioral goals and users' percep-

[1]Technologies have the same level of tools, methodologies, techniques, among others (Petersen et al., 2015).

Guerino, G., Silva, W., Coleti, T. and Valentim, N.
Assessing a Technology for Usability and User Experience Evaluation of Conversational Systems: An Exploratory Study.
DOI: 10.5220/0010450204630473
In *Proceedings of the 23rd International Conference on Enterprise Information Systems (ICEIS 2021) - Volume 2*, pages 463-473
ISBN: 978-989-758-509-8

tions.

To a deeper understanding of software quality evaluation, we conducted two systematic mappings studies, and we realized that technologies used to evaluate the Usability and UX of CSs are generical and do not include specific CS aspects, e.g., the conversation's rhythm. Besides, some technologies evaluate only one quality aspect (Usability or UX) in CSs.

Thus, this paper presents a proposal for a joint evaluation of Usability and UX aspects specific to CSs, the Usability and User Experience Evaluation of Conversational Systems (U2XECS). U2XECS aims to help developers how to evaluate their CSs, identify improvement points, and understand users' perceptions. Besides, UEXECS was created to be used by developers in an easy and low-cost way. This paper also presents an exploratory study to improvement and evolve the U2XECS. To guide our study, we designed the following research questions:

- RQ1. How efficient were participants using U2XECS?
- RQ2. What the participants' perceptions of U2XECS based on ease of use, usefulness, and intention to use?
- RQ3. Which are the improvements suggested for U2XECS?

To answer the presented questions, the study verified participants' efficiency in identifying defects in the CS evaluated through U2XECS (RQ1). The study also evaluated the U2XECS through the Technology Acceptance Model (TAM) (Venkatesh and Davis, 2000) according to ease of use, usefulness, and intention to use (RQ2). Besides, we proposed one qualitative question to identify possible improvements to our technology (RQ3).

This paper is organized as follows: Section 2 presents related works. Section 3 presents the U2XECS, its definition, mode and example of use. Section 4 presents the planning and execution of the exploratory study. In Section 5, we present the exploratory study results. In Section 6, we present the discussion, improvements identified in the study, and the updated version of U2XECS. In Section 7, we discuss the threats to validity. Finally, Section 8 presents our conclusions and future work.

2 RELATED WORK

The literature reveals that CSs are evaluated in different ways by different technologies (Kocaballi et al., 2018). Technologies identified in the literature and that are commonly employed to evaluate the Usability and/or UX of CSs can be of three types: generic, i.e., they can be used to evaluate any system; developed by the authors themselves for a specific context or study; they can be specific for CSs. We present below some studies in which the authors used some evaluation technologies.

Some generic technologies are used in the CS context. System Usability Scale (SUS) (Brooke, 1996) is a ten-item questionnaire that can be answered using a 5-point Likert scale to evaluate the ease of use and learnability of the system. Another example is AttrakDiff (Hassenzahl et al., 2003), a technology composed of twenty-eight items to evaluate the pragmatic quality, hedonic quality, and attractiveness of a system. For the CS context, they can be limited to evaluate a CS's specificity, such as the conversation's rhythm.

We also found studies developing their evaluation questionnaires (Di Nuovo et al., 2018; Lee et al., 2017). On the one hand, creating questionnaires allow evaluating aspects that the authors need for that study. On the other hand, these developed questionnaires do not undergo a process of empirical validation. Lazar et al. (2017) state that experimental research is a relevant approach to making systematic judgments about a technology's reliability.

Hone and Graham (2000) developed the SASSI (Subjective Assessment of Speech System Interfaces) questionnaire. SASSI consists of 34 items divided into six categories: system response accuracy, likeability, cognitive demand, annoyance, habitability, and speed (Hone and Graham, 2000). The answers available for the items are based on a 7-point Likert scale. Despite the importance of this technology for the CS context, it does not consider aspects related to user perceptions. Moreover, we have not found an up-to-date version of SASSI since this technology was published in 2000, and in recent years CSs have undergone several modifications.

One UX-specific evaluation technology for the CS context is SUXES (User Experience Evaluation Method for Spoken and Multimodal Interaction) (Turunen et al., 2009). SUXES is a four-stage evaluation procedure: (i) introduction to the evaluation, background questionnaire, and time reservation for the test; (ii) introduction to the application and expectation questionnaire; (iii) user experiment and experience questionnaire; and (iv) opinion questionnaire. However, despite the relevant methodological basis, the authors do not provide all the questionnaires' statements. Besides, usability aspects were not considered for evaluation in this technology.

Although some technologies cited above are

widely used in the literature, we have not identified any efforts made to propose a technology specific to the current CS context, and that evaluates Usability and UX jointly. Thus, our paper focuses on describing a novel technology that can fill these gaps: specific to CS context and that evaluate Usability and UX jointly.

3 U2XECS

This section presents the definition of U2XECS and how it can be used in the evaluation process of a CS.

3.1 Definition of U2XECS

First, we performed a comprehensive Systematic Mapping Study (SMS) (Guerino and Valentim, 2020b) for non-conventional interactions. We identified 30 different Usability or UX evaluation technologies that are used in non-conventional interactions. However, we concluded that these technologies are generic (they can evaluate any system) and evaluate only one quality criterion, Usability, or UX.

Then, we decided to conduct another SMS (Guerino and Valentim, 2020a), more specific to the CSs context. We identified 31 Usability or UX evaluation technologies that are being used in the context of CS. The analysis of these technologies revealed that researchers usually create their evaluation questionnaires for their specific studies. These questionnaires generally do not go through an empirical evaluation, which can compromise and bias the results.

Based on the gaps found in the SMSs results, we proposed the U2XECS, a questionnaire-based technology to evaluate Usability and UX jointly in CSs. To define the U2XECS statements, we used aspects based on ISO 9241-210 (2019) to classify the Usability statements: efficiency, effectiveness, and user satisfaction. We also used dimensions proposed by Bargas-Ávila and Hornbæk (2011): generic UX, affect/emotion, enjoyment/fun, aesthetics/appeal, engagement/flow, motivation, and enchantment. We chose these aspects mentioned above because they are consolidated works in literature. U2XECS consists of statements that can be answered by a 5-points Likert scale and qualitative statements that can be answered by text. Table 1 presents the aspects and some statements of U2XECS. The complete version of U2XECS is available in https://bit.ly/2FecMth.

3.2 Mode and Example of Use

U2XECS can be employed when developers need to evaluate CSs with their end-users. Figure 1 illustrates how U2XECS can assist in the CS evaluation. The steps to use U2XECS are: (i) participants perform tasks in the CS to be evaluated; (ii) participants respond to U2XECS about the CS they used; (iii) developers analyze the responses provided by users to identify the points to be improved in their CS. We consider that, from these steps, the CS can have a more friendly version that provides a relevant experience to its users.

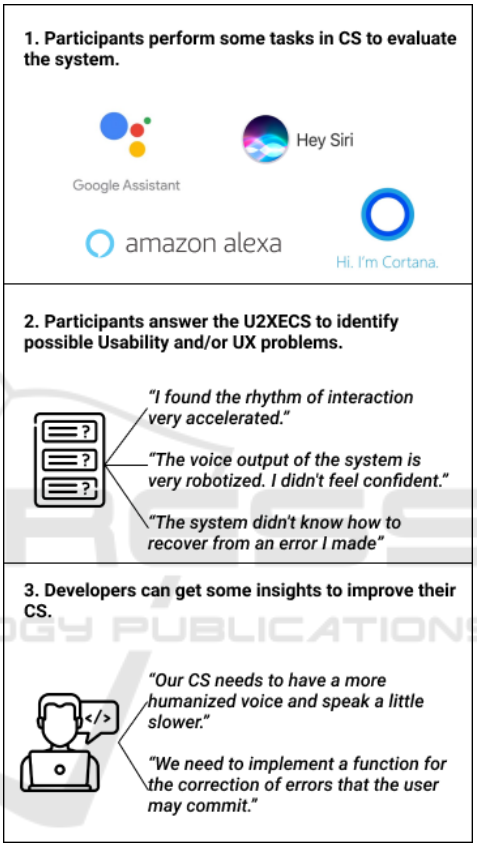

Figure 1: Example of U2XECS use.

4 EXPLORATORY STUDY

We performed an exploratory study aiming to evaluate the participants' efficiency when identifying defects in CS with U2XECS, whether participants considered U2XECS easy to use and useful, and they intend to use it in the future. Moreover, we intend to identify possible improvements for U2XECS.

4.1 Context

In this exploratory study, we selected the Amazon Alexa application to be the object study evaluated by U2XECS. Amazon, with its virtual assistant Alexa,

Table 1: Aspects and some statements of U2XECS.

Aspects	**# Item/Statement**
User Satisfaction	2. I needed to learn a lot about the system before performing these tasks with my voice.
Efficiency	10. The rhythm of voice interaction with the system was appropriate.
Effectiveness	15. The conversational system was able to recover easily from some error or mistake I made.
Generic UX	21. Performing these tasks with the voice in the system was a good experience.
Affect/Emotion	25. The system answered my interaction by voice in a friendly way.
Enjoyment/Fun	26. It was very pleasant to use voice to perform these tasks in the system.
Aesthetics/Appeal	29. The system had an innovative design that made it easier to perform tasks through voice.
Engagement/Flow	33. I felt in control of the system during the voice interaction.
Motivation	34. I felt motivated when using voice to perform these tasks in the system.
Enchantment	36. The tasks I did in the system with my voice made me enchanted with this kind of interaction.
Qualitative	38. A positive point when using voice interaction to perform these tasks is: _____, because: _____

is at the top of the smart speaker market since 2017, and the trend is to continue in 2021 (eMarketer, 2020). Only in the Play Store (Android devices), the Amazon Alexa application was downloaded by over 50 million users.

4.2 Participants

Thirty-three students of HCI and Software Engineering (SE) courses from three different universities participated in the study. Participants filled out a characterization form to categorize their expertise. For system development and CS development, we considered: (i) high experience, participants who had worked in more than five systems/CS development projects in the industry; (ii) medium experience, participants who had worked from 1 to 4 system/CS development projects in industry; (iii) low experience, participants who worked in at least one system/CS development project in the industry; and (iv) with no experience, participants who had no previous knowledge system/CS development project or who had some knowledge but no practical experience. For Usability/UX evaluation, we considered: (a) high experience, participants who had worked in Usability/UX projects/evaluations in the industry; (b) medium experience, participants who had worked in Usability/UX projects/evaluations in the classroom; (c) low experience, participants who had knowledge of Usability/UX acquired in lectures or readings; and (d) with no experience, participants who had no knowledge about Usability/UX. Table 2 (second, third, and fourth column) demonstrates each participants' categorization. The label 'P' and a number identify each participant, e.g., P1 identifies participant 1.

4.3 Indicators

We defined four indicators: efficiency, ease of use, usefulness, and intention to use.

- Efficiency: a total of defects identified by the participant divided by the total time spent using U2XECS to evaluate the CS.

After evaluating the Amazon Alexa application using U2XECS, the participants answered the post-study questionnaire based on TAM indicators (Venkatesh and Davis, 2000) about using U2XECS. TAM is a widely used model for assessing the acceptance of new technologies (Marangunić and Granić, 2015). The TAM has three indicators, which are:

- *Perceived Ease of Use.* Defines the degree to which a person believes that using a specific technology would be effortless through the following statements: (E1) My interaction with the U2XECS is clear and understandable, (E2) Interacting with the U2XECS does not require a lot of my mental effort, (E3) I find the U2XECS to be easy to use, and (E4) I find it easy to get the U2XECS to do what I want it to do, to evaluate the Usability and UX of CS.
- *Perceived Usefulness.* Defines the degree to which a person believes that technology could improve their performance through the following statements: (U1) Using the U2XECS improves my performance in the evaluation of Usability and UX of CS, (U2) Using the U2XECS in the evaluation of Usability and UX of CS increases my productivity, (U3) Using the U2XECS enhances my effectiveness in the evaluation of Usability and UX of CS and (U4) I find the U2XECS to be useful in the evaluation of Usability and UX of CS.

- *Intention to Use.* Defines the degree to which a person believes they would use the technology in future projects through the following statements: (I1) Assuming I have access to the U2XECS, I intend to use it and (I2) Given that I have access to the U2XECS, I predict that I would use it.

We also used one open-ended question to extract further improvements for U2XECS: *In your opinion, how could the U2XECS be improved?*

4.4 Instrumentation

We used some artifacts (all available on Google Forms) to support the exploratory study: (i) a consent form[2], to ensure the confidentiality and privacy of the data collected; (ii) the characterization form[3], which had questions to characterize the expertise about the development and evaluation of systems; (iii) the instruction sheet[4], containing the tasks to be performed by participants in the study; (iv) the U2XECS[5]; and (v) a post-study questionnaire[6] containing TAM indicators and one open-ended question.

4.5 Preparation

Before execution, we conducted training with the participants. We presented to them about Usability and UX evaluation, CSs, and U2XECS. We performed a 30-minute presentation on the topics mentioned. Examples of CSs, Usability/UX evaluation technologies, and the U2XECS were provided. At the end of each presentation, the authors answered doubts from the participants about the concepts covered.

4.6 Execution

Due to the pandemic caused by COVID-19, the study had to be adapted for online context. We conducted a meeting for each class by Google Meet[7]. After preparing the participants, we sent the instructions where we describe the activities and tasks to be performed. The participants executed the activities and tasks and sent an instruction document containing the name and time spent on tasks. The answers to the other artifacts were all recorded by Google Forms. The participants performed the following steps in the study:

- STEP 1
  - Participants agreed to the consent form and answered the characterization questionnaire;
- STEP 2
  - By cell phone, the participants downloaded and configured the Amazon Alexa application;
  - Participants noted the initial time;
  - Participants did the following tasks in the Amazon Alexa application:
    - Ask Alexa: “Alexa, what things can I try?”
    - Ask Alexa when is the independence day;
    - Ask Alexa to spell the word “Pumpkin”;
    - Ask Alexa what the word “Measurement” means;
    - Ask Alexa who is “Elon Musk”;
    - Ask Alexa the weather forecast in her city;
    - Ask Alexa about the weather in Curitiba;
    - Ask Alexa how many days are left for Christmas;
    - Ask Alexa how much is 756 divided by 9;
    - Ask Alexa how much Brazilian money is 25 dollars.
  - After finishing the tasks above, participants evaluated the Amazon Alexa application through U2XECS;
  - After answering the U2XECS, the participant noted the final time;
- STEP 3
  - After using the U2XECS, participant evaluated U2XECS about its ease of use, usefulness, and intention to use, and answered the open-ended question;

5 RESULTS

5.1 RQ1. Participants’ Efficiency

Table 2 shows the overall results of the participants’ evaluation. We verified that participants identified between 0 and 9 defects in the Amazon Alexa application, and they spend between 9 and 70 minutes. P4 obtained greater efficiency compared to other participants, identifying about 16.80 defects per hour. We noticed that most of the participants who had the highest efficiency (P2, P4, P10, and P24) had medium experience with Usability and UX evaluation, which may have helped identify defects. On the other hand, most participants with low efficiency (P3, P7, P14, P17, and P23) had none or low experience with Usability and UX evaluation, which may have influenced the identification of defects in the application.

The researchers analyzed all the participants’ responses to consider defects or not. Considering duplicates, participants described 100 defects that they identified in the Amazon Alexa application through

[2]Consent Form: https://bit.ly/3bMEHwg
[3]Characterization Form: https://bit.ly/2ZvCZdT
[4]Instruction Sheet: https://bit.ly/3k9cYc8
[5]U2XECS: https://bit.ly/2FecMth
[6]Post-study Questionnaire: https://bit.ly/33eyw0g
[7]https://meet.google.com/

Table 2: Summary of characterization and identification of defects by participants.

Partic.	SD	CSD	UUXE	# Disc.	# False Pos.	# Defects	Time (min)	Defects/Hour
P1	Low	None	Low	4	1	3	21	8.57
P2	Medium	None	Medium	10	1	9	35	15.43
P3	Low	None	Low	0	0	0	30	0.00
P4	None	None	None	7	0	7	25	16.80
P5	None	None	None	5	0	5	23	13.04
P6	Low	None	Low	4	0	4	22	10.91
P7	Medium	None	Low	1	0	1	29	2.07
P8	Medium	None	Low	2	0	2	52	2.31
P9	Medium	Low	None	4	1	3	31	5.81
P10	Low	None	Medium	6	0	6	24	15.00
P11	None	None	Medium	0	0	0	10	0.00
P12	None	None	None	3	0	3	67	2.69
P13	Low	None	Medium	3	0	3	70	2.57
P14	None	None	Low	2	0	2	62	1.94
P15	Low	Medium	Medium	6	0	6	46	7.83
P16	None	Low	Low	4	0	4	22	10.91
P17	None	None	Low	1	1	0	15	0.00
P18	None	None	Medium	6	1	4	31	7.74
P19	Low	Low	Medium	3	0	3	46	3.91
P20	Low	None	Low	1	0	1	23	2.61
P21	Low	None	Medium	2	0	2	56	2.14
P22	None	None	None	1	0	1	14	4.29
P23	None	None	None	2	1	1	29	2.07
P24	Low	None	Medium	5	1	4	15	16.00
P25	Low	Low	Medium	4	0	4	31	7.74
P26	Low	None	None	4	1	3	25	7.20
P27	Low	None	Medium	1	0	1	29	2.07
P28	Low	None	Medium	2	0	2	9	13.33
P29	Low	None	Low	3	0	3	11	16.36
P30	Low	None	Medium	2	0	2	20	6.00
P31	Low	None	Medium	3	1	2	48	2.50
P32	Low	None	Medium	6	1	5	46	6.52
P33	Low	None	Medium	4	0	4	44	5.45

Legend - **SD**: Experience in System Development; **CSD**: Experience in Conversational System Development; **UUXE**: Experience in Usability/User Experience Evaluation; **Disc**: Number of discrepancies; **False Pos**: Number of false positives.

U2XECS. Without considering the duplicates, 48 different defects were identified. Examples of these defects are "I did not feel so confident because of the wrong answer the assistant gave", "Alexa did not return suggestions for correcting the word she did not identify", and "There are many icons that can make learning difficult". Besides, we have observed that students have identified few false positives (Table 2).

5.2 RQ2. Participants' Perceptions based on Ease of Use, Usefulness, and Intention to Use

The analysis of participants' perception of this study was based on the TAM statements. Figure 2 illustrates the overall results obtained in this study.

Perceived Ease of Use. The results indicate that participants realized the U2XECS is easy to use. Regarding E1, 88% of participants (N = 29) agreed or partially agreed that interaction with the U2XECS was clear and understandable. Related to E2, 58% of participants (N = 19) agreed or partially agreed that U2XECS did not require a lot of mental effort. Still, in relation to E3, 79% of participants (N = 26) agreed or partially agreed that U2XECS was easy to use. Moreover, in E4, 85% of participants (N = 28) agreed or partially agreed that they found the U2XECS easy to use for what it has to do, to evaluate the Usability and UX of a CS.

However, we verified that in E2, some participants did not provide a positive response to the U2XECS. About 42% of participants (N = 14) did not agree or disagree, or partially disagreed or even totally disagreed that U2XECS did not require a lot of mental effort. This result indicates that a relevant mental effort was required for several participants to respond to the U2XECS. The size of the U2XECS could be a factor that influenced this mental effort request.

Perceived Usefulness. The TAM results indicate that

Figure 2: Overall results on U2XECS obtained with TAM.

the participants realized the usefulness of U2XECS but have not yet convinced of it. Regarding U1, a balance was identified in the participants' opinions. About 52% of participants (N = 17) agreed or partially agreed that U2XECS could improve their performance in evaluating the Usability and UX of CS. However, 48% of participants (N = 16) neither agreed nor disagreed about the performance increase. Related to E2, positive results were more present. About 70% of participants (N = 23) agreed or partially agreed that using U2XECS can increase productivity. In relation to E3, we noticed a balance between the responses. About 61% of participants (N = 20) agreed or partially agreed that U2XECS could enhance the effectiveness of the evaluation of Usability and UX of CS. On the other hand, 39% of participants (N = 13) did not agree or disagree. Finally, at U4, about 82% of participants (N = 27) consider U2XECS useful in evaluating the Usability and UX of CS.

Negative responses regarding the usefulness of U2XECS may be related to the experience of our participants. As demonstrated in Table 2, some participants had none or low experience with Usability or UX evaluation. Moreover, most participants had none experience with CS development. Due to these factors, some participants did not find U2XECS useful.

Intention to Use. The results of the TAM indicate that participants intend to use U2XECS in the future. However, we also noticed that some participants were unsure about the future use of U2XECS. Regarding I1, 67% of participants (N = 22) agreed or partially agreed that having access to the U2XECS, they would use the technology. On the other hand, 33% of participants (N = 11) neither agreed nor disagreed, partially disagreed, or even totally disagreed with the technology's future use. Related to I2, 64% of participants (N = 21) agreed or partially agreed that having access to the U2XECS, they predict that they would use it on other occasions. However, 36% of the participants (N = 12) neither agreed nor disagreed, or partially disagreed, or even totally disagreed with the prediction of using U2XECS.

The same factors that influenced the easy of use and usefulness indicator also influenced the intentions to use. The low experience, especially with CS, may have influenced the negative responses since some participants have never participated in CS development and have no intention to participate, influencing the opinion about the intention to use.

5.3 RQ3. Improvements Suggested

The qualitative analysis was based on the reading of all open-ended responses. We classified the results into three categories: questionnaire benefits, negative points, and suggestions for improvement.

5.3.1 Questionnaire Benefits

The participants enjoyed the U2XECS and its characteristics, such as the composition of the statements, usefulness, and structure. Participants liked the clarity of the statements and the format they are displayed (see P5 quotation). They also identified some useful points of the U2XECS and how it can support the researcher using it (see P9 quotation). Moreover, the structural division between Usability and UX pleased the participants (see P27 quotation).

- *"As a benefit, the U2XECS makes clear the statements and what it is analyzing."* [P5]

- *"The U2XECS made me questions about points of the application that sometimes I ignore without wanting to."* [P9]
- *"I had no difficulties with the U2XECS; the statements were very clear and well-grouped in specific topics."* [P27]

5.3.2 Negative Points

The negative points that the participants identified in the U2XECS were about the questionnaire's size and difficulty in understanding some statements. Participants found the questionnaire relatively long (see P1 quotation), which left them tired (see P8 quotation). Besides, participants mentioned the difficulty of understanding some statements (see P4 quotation) and the repetition of some of them (see P10 quotation).

- *"The questionnaire is very long, which makes it tiring near the end."* [P1]
- *"The questionnaire is too big; it made me bored."* [P8]
- *"The U2XECS is easy to use but tiring since the texts are long and sometimes difficult to understand."* [P4]
- *"It is a bit extensive, and some statements are a bit similar."* [P10]

5.3.3 Suggestions for Improvement

Participants contributed by suggesting improvements to the U2XECS regarding functionality, response types, and overall structure. Some of these suggestions were implemented in the updated version of U2XECS (see Appendix A) and is discussed in the Discussion Section. Other suggestions will be better analyzed for future implementation. Participants suggested that statements could be answered by voice (see P5 quotation). Finally, participants suggested reducing the textual responses (see P32 quotation) and mentioned that each aspect's qualitative statement should be optional (see P8 quotation).

- *"U2XECS could implement voice evaluation features, such as answering a question through speech, which would help shorten filling time."* [P5]
- *"Elaborate statements that can replace statements that require textual answers (reduce textual answers) but maintain the questionnaire's effectiveness."* [P32]
- *"The justification at the end of each section of the aspect should be optional."* [P8]

6 DISCUSSION AND IMPROVEMENT OF U2XECS

The study has contributed to the identification of improvements to our technology. U2XECS has gone through an evolutionary process and can be seen in Appendix A. In this section, we will discuss the study results and how they have influenced our decisions.

First, we realized that qualitative statements at the end of U2XECS may leave their use boring since the participant requires much time to read all the statements and still has to write about them. Besides, qualitative statements at the end may generate repeated defects, i.e., defects that have already been identified along the U2XECS. Therefore, we have decided to remove the qualitative statements at the end of U2XECS and leave them only at the end of each set of statements by the Usability/UX aspect.

We identified that the most significant negative points were the questionnaire's size and similar statements. Therefore, we decided to remove some statements from the U2XECS to address these identified suggestions. The statement withdrawal process was based on three choices: (i) statements from aspects that identified false positives; (ii) statements that did not directly contribute to the identification of defects; and (iii) statements that could be similar.

We have noticed that statements of motivation and enchantment have contributed to the identification of false positives. The false positives identified with the U2XECS were the considerations that the participants made but did not characterize a defect. For instance, in response to the enchantment aspect, a participant answered: "*I would be more enchanted if I had a real-time transcription of the speech*". This contribution is not a defect because the lack of real-time transcription is not a negative aspect of the system. Therefore, it was characterized as a false positive. We have decided to withdraw some statements regarding motivation and enchantment aspects, which are: "*I think I would use voice interaction regularly in this system instead of any other kind of interaction*" and "*The tasks I did in the system with the voice made me enchanted with this kind of interaction*". We realized that they do not encourage participants to judge the CS because they depend on the user's preference to use the voice to perform the actions.

Then, we reread all the remaining statements in the U2XECS to reduce the questionnaire's size. We also identified another question that we decided to remove, this time in the generic UX aspect: "*I have never had a voice interaction experience like I did when performing these tasks in the system*". This question was removed because it does not judge the CS being evaluated in the study but compared the use with the participant's experience.

We also analyzed the statements that could be similar. We excluded the Enjoyment/Fun aspect: "*It was very boring to use voice to perform these tasks in the*

system". This question can be considered the opposite of the question about pleasure: "*It was very pleasant to use voice to perform these tasks in the system*". In other words, the participant who found the interaction boring, by logic, did not find it pleasant.

We also identified statements that could generate ambiguity in the participants' opinions. One of them was from the efficiency aspect: "*During my voice interaction with the system, I had to repeat several commands*". The other, from the effectiveness aspect: "*The voice interaction system forced me to use keywords*". These statements were removed because CS, by nature, uses keywords to be triggered (e.g., "Hey, Siri!" and "Ok, Google"). However, this is not a defect of this type of systems, but a CS characteristic.

Summarizing, the improvements for the current version of U2XECS were: (i) removal of qualitative statements at the end of U2XECS, leaving only at the end of each aspect; (ii) removal of two statements that identified false positives; (iii) removal of one statement that did not judge the CS approved, but rather the previous experiences; (iv) removal of one statement that could be repeating the meaning of others; and (v) removal of two statements that could generate ambiguous opinions in participants.

7 THREATS TO VALIDITY

Some threats could affect our study validity, and we discussed them in this Section. We have classified the threats according to Wöhlin et al. (Wohlin et al., 2012): internal, external, conclusion, and construct.

Internal Validity. In our study, we considered two main threats that risk for an inadequate interpretation of the results: (1) the effects of training and (2) the classification of experience. There could be a training effect if U2XECS training was conducted differently depending on the class. We control the training effects by using the same material for all classes and removing all doubts as they arise. Regarding the participant's experience, this was based on the self-classification. They self-classified according to the number and type of previous experiences with system and CS development, and Usability/UX evaluation.

External Validity. One threat was considered, the participants were undergraduate students, and few participants had experience in the industry. Carver et al. (Carver et al., 2004) state that students who do not have industry experience may have similar skills to less experienced evaluators.

Conclusion Validity. Another threat is that the study was conducted remotely, obeying the social isolation caused by the COVID-19 pandemic. We could not control the bias caused by external factors. Some elements outside the scenario may have disturbed the results, e.g., a noise outside the room or a interruption during the study's execution. However, based on the results, we consider that participants fulfilled all the tasks and contributed to our technology's evolution.

Construct Validity. Two threats were considered: (1) guesswork of goals and results, and (2) apprehension of evaluation. Regarding (1), the participants could try to determine the goal and the result intended by the researchers. They could base their assumptions about the results, positively or negatively, depending on their attitude towards the anticipated goals. To reduce this threat, participants were told several times that there were no answers expected by the researchers and that we wanted to evaluate our technology. Regarding (2), some participants may be afraid of being evaluated. One way of human tendency is to try to look better when evaluated, which can bias the results. To mitigate this threat, we explained to the participants that there was no right or wrong answer and that the important was their opinion.

8 CONCLUSION

We presented the U2XECS technology and a study conducted in the process of proposal and evolution of our technology. The quantitative results of the TAM indicators pointed out that most participants found U2XECS easy to use and useful and intended to use it in the future. From the qualitative analysis, we checked the faulty points of U2XECS, mainly related to the questionnaire's size. Therefore, we proposed the updated version of U2XECS, a smaller and optimized version, without losing the specificity for CS.

Overall, U2XECS contributions are: (a) can be applied in evaluations with potential users before making the CS available to society; (b) assists developers in identifying defects in CSs based on Usability and UX aspects; (c) the development team can focus on solving the real problems identified in the evaluation; (d) the time to release the product to market decreases, as Usability and UX problems are identified and corrected during the CS development.

Two studies with industry professionals will guide future works. First, we will make our technology available for the professionals to inspect. Then we will interview each one to verify their opinions. Second, we will evaluate the U2XECS in a real context of use, assisting a group of professionals developing and evaluating a CS. We hope that U2XECS can assist CS developers in their evaluations with potential users.

ACKNOWLEDGEMENTS

We thank all study participants who contributed to the evolution of our technology.

REFERENCES

Brooke, J. (1996). A "Quick and Dirty" Usability Scale. In Jordan, P. W. and et al., editors, *Usability Evaluation in Industry*, pages 189–194. London: Taylor & Francis.

Candello, H., Pinhanez, C., Pichiliani, M., Cavalin, P., Figueiredo, F., Vasconcelos, M., and Do Carmo, H. (2019). The effect of audiences on the user experience with conversational interfaces in physical spaces. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–13. ACM.

Carver, J., Jaccheri, L., Morasca, S., and Shull, F. (2004). Issues in using students in empirical studies in software engineering education. In *Proceedings of 5th International Workshop on Enterprise Networking and Computing in Healthcare Industry*, pages 239–249. IEEE.

Di Nuovo, A., Broz, F., Wang, N., Belpaeme, T., Cangelosi, A., Jones, R., Esposito, R., Cavallo, F., and Dario, P. (2018). The multi-modal interface of robot-era multi-robot services tailored for the elderly. *Intelligent Service Robotics*, 11(1):109–126.

eMarketer (2020). Amazon maintains convincing lead in us smart speaker market. https://www.emarketer.com/content/amazon-maintains-convincing-lead-in-us-smart-speaker-market.

Grand View Research, Inc (2020). Intelligent virtual assistant market size, share & trends analysis report by product (chatbot, smart speakers), by technology, by application (bfsi, healthcare, education), by region, and segment forecasts, 2020 - 2027. https://www.grandviewresearch.com/press-release/global-intelligent-virtual-assistant-industry.

Guerino, G. C. and Valentim, N. M. C. (2020a). Usability and user experience evaluation of conversational systems: A systematic mapping study. In *Proceedings of the 34th Brazilian Symposium on Software Engineering*, SBES '20, page 427–436. ACM.

Guerino, G. C. and Valentim, N. M. C. (2020b). Usability and user experience evaluation of natural user interfaces: a systematic mapping study. *IET Software*, 14(5):451–467.

Hassenzahl, M., Burmester, M., and Koller, F. (2003). Attrakdiff: Ein fragebogen zur messung wahrgenommener hedonischer und pragmatischer qualität. In *Mensch & Computer*, pages 187–196. Vieweg+Teubner Verlag.

Hone, K. and Graham, R. (2000). Towards a tool for the subjective assessment of speech system interfaces (sassi). *Natural Language Engineering*, 6(3-4):287–303.

Kocaballi, A. B., Coiera, E., and Berkovsky, S. (2020). Revisiting habitability in conversational systems. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI EA '20, page 1–8. ACM.

Kocaballi, A. B., Laranjo, L., and Coiera, E. (2018). Measuring User Experience in Conversational Interfaces: A Comparison of Six Questionnaires. In *Proceedings of the 32nd International British Computer Society Human Computer Interaction Conference*, pages 1–12, Belfast, Northern Ireland. BCS Learning & Development Ltd.

Lee, J., Lee, C., and Kim, G. J. (2017). Vouch: Multimodal touch-and-voice input for smart watches under difficult operating conditions. *Journal on Multimodal User Interfaces*, 11(3):289–299.

Lovato, S. and Piper, A. M. (2015). "siri, is this you?": Understanding young children's interactions with voice input systems. In *Proceedings of the 14th International Conference on Interaction Design and Children*, IDC '15, page 335–338. ACM.

Marangunić, N. and Granić, A. (2015). Technology acceptance model: a literature review from 1986 to 2013. *Universal Access in the Information Society*, 14:81–95.

Petersen, K., Vakkalanka, S., and Kuzniarz, L. (2015). Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology*, 64:1–18.

Porcheron, M., Fischer, J. E., Reeves, S., and Sharples, S. (2018). Voice interfaces in everyday life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–12. ACM.

Trajkova, M. and Martin-Hammond, A. (2020). "alexa is a toy": Exploring older adults' reasons for using, limiting, and abandoning echo. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13. ACM.

Turunen, M., Hakulinen, J., Melto, A., Heimonen, T., Laivo, T., and Hella, J. (2009). SUXES - user experience evaluation method for spoken and multimodal interaction. In *INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association, Brighton, United Kingdom*, pages 2567–2570. ISCA.

Venkatesh, V. and Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science*, 46(2):186–204.

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., and Wesslén, A. (2012). *Experimentation in Software Engineering*. Springer-Verlag Berlin Heidelberg.

APPENDIX A

Updated Version of U2XECS

Table 3: Aspects and statements of the updated version of U2XECS.

User Satisfaction					
	1	**2**	**3**	**4**	**5**
1. It was easy to use voice to perform the tasks in this system.					
2. I did not need to learn a lot about the system before performing these tasks with my voice.					
3. I was able to familiarize myself with the system when I first used it.					
4. I felt satisfied using my voice to perform these tasks.					
5. The system behaved the way I expected during the voice interaction.					
6. I found it easy to understand how to interact by voice in the system.					
7. It was easy to become skilled when using the system.					
* Based on the statements above and your answers, describe the satisfaction issues you identified in the system.					
Efficiency					
8. The system immediately responded to my voice interaction.					
9. I would frequently use voice to perform these tasks in the system.					
10. The rhythm of voice interaction with the system was appropriate.					
11. I was able to complete my tasks with voice interaction in a time that I believe is reasonable.					
* Based on the statements above and your answers, describe the efficiency issues you identified in the system.					
Effectiveness					
12. The system recognizes what I said during my voice interaction.					
13. The system was able to recover quickly from some error or mistake I made.					
15. When the data entered in the system was inconsistent or ambiguous, the system requested more information.					
16. From the novice users' point of view, the system has led the interaction by voice in the right way.					
16. From the novice users' point of view, the system has led the interaction by voice in the right way.					
17. For advanced users, the system allowed a large amount of input data at once.					
18. The speed of my internet does not influence the system outcomes.					
* Based on the statements above and your answers, describe the effectiveness issues you identified in the system.					
Generic UX					
19. Performing these tasks with the voice in the system was a good experience.					
20. The system can handle my accent and language characteristics.					
* Based on the statements above and your answers, describe the generic UX issues you identified in the system.					
Affect/Emotion					
21. I felt discouraged when using voice to perform these tasks in the system.					
22. I did not feel confident when using voice to interact with this system.					
23. The system answered my interaction by voice in a friendly way.					
* Based on the statements above and your answers, describe the affect and emotion issues you identified in the system.					
Enjoyment/Fun					
24. It was very pleasant to use voice to perform these tasks in the system.					
25. I had fun using my voice to perform these tasks in the system.					
* Based on the statements above and your answers, describe the enjoyment and fun issues you identified in the system.					
Aesthetics/Appeal					
26. The system had an innovative design that made it easier to perform tasks through voice.					
27. Using voice to perform the tasks attracted me to use the system.					
* Based on the statements above and your answers, describe the aesthetics/appeal issues you identified in the system.					
Engagement/Flow					
28. I concentrate on doing the tasks when using voice interaction.					
29. I felt in control of the system during the voice interaction.					
* Based on the statements above and your answers, describe the engagement/flow issues you identified in the system.					
Motivation					
30. I felt motivated when using voice to perform these tasks in the system.					
* Based on the statements above and your answers, describe the motivation issues you identified in the system.					