

In Search of a Conversational User Interface for Personal Health Assistance

Mathias Wolfgang Jesse¹^a, Claudia Steinberger²^b and Peter Schartner²^c

¹Digital Age Research Center, University of Klagenfurt, Universitätsstraße 65-67, Klagenfurt, Austria

²Department of Applied Informatics, University of Klagenfurt, Universitätsstraße 65-67, Klagenfurt, Austria

Keywords: Health Technologies, Innovative Interfaces, Conversational User Interface, Voice Assistant, Ehealth, AAL.

Abstract: Conversational user interfaces (CUI) cause a paradigm shift in the interaction between user and machine. The machine is operated via structured dialogues or partly or entirely via human language. Voice assistants that understand and process spoken natural language are increasingly being used. The currently available conversational technologies (CTs) for voice assistants range from well-known commercial technologies to quite well-known open source platforms. The suitability of a CT for a particular application domain depends heavily on its specific requirements. In this paper, we focus on the selection of CTs for the development of CUIs for the elderly to assist them in their health management. We (1) propose criteria for CT selection in the domain of personal health management for the elderly, (2) analyze commercial and open source representatives according to these criteria and (3) we evaluate the most suitable candidates for CUI development.

1 INTRODUCTION


Language is one of the most powerful forms of communication between people. Technological developments in recent years, especially in the field of artificial intelligence (AI), natural language processing (NLP) and pervasive computing, have opened new possibilities to realize human-to-computer interfaces. New functions and tools have emerged that are essential for computer-aided processing of spoken natural language. Conversational user interfaces (CUIs) caused a paradigm shift in the human-to-computer interaction (Hoy, 2018; Kěpuska & Bohouta, 2018; Siddike et al., 2018).


Particularly intensive work was done on technologies to understand and process spoken language (Kinsella & Mutchle, 2019; Olson & Kemery, 2019; Siddike et al., 2018; SUMO Heavy Industries, 2019). Currently available conversational technologies (CTs) for voice assistant development and deployment span from familiar commercial representatives like Apple Siri, Amazon's Alexa, Google Assistant or Microsoft Cortana to quite well-known open source technologies like Rhasspy or Mycroft. These CTs


have developed rapidly in recent years and they offer different development possibilities and features. Nowadays, one can use these technologies to extend existing or build own voice assistants and offer these functionalities on various platforms.

Voice assistants can be helpful to the user in different ways. The users who benefit most are those who are restricted to use a graphical user interface (GUI). This group includes disabled persons or persons in special working or living situations, users with low digital skills or users simply preferring to use their voice to control applications. Furthermore, the use of voice enables novice users to directly interact with a device without having to adapt to a GUI. More experienced users can utilize a CUI for a faster and more direct interaction (Huang et al., 2001).

In addition, in comparison to a GUI the interaction with a CUI can be made more dynamic and flexible, adapting better to the cognitive abilities of the user (Siddike et al., 2018; Wolters et al., 2015). However, not every technology is equally suitable for every problem. The appropriateness of a CT depends on the requirements of an application domain. The scope of

^a <https://orcid.org/0000-0001-6018-2780>

^b <https://orcid.org/0000-0002-5111-2286>

^c <https://orcid.org/0000-0002-5964-8480>

this paper lies within the scope of active assisted living (AAL) support for the elderly. The focus is on analyzing available CTs to develop CUIs for personal health assistance for the elderly. The paper investigates the applicability of CTs in this domain with special requirements e.g. on data protection and security, multimodality and special needs of the target group.

The contribution of this paper is to (a) propose criteria for the evaluation of CTs for the eHealth application domain, to (b) analyze representatives of commercial CTs, open source CTs and open source text-to-speech and voice recognition components available on the market to these criteria and (c) to evaluate their suitability to develop and deploy multimodal solutions for the personal eHealth management for the elderly.

The paper is structured as follows: Chapter 2 introduces the research methodology applied in this work. Chapter 3 provides an overview of the state of art in CTs. Chapter 4 outlines our evaluation criteria and explains how they were determined. Chapter 5 selects technologies for further evaluation. Chapter 6 evaluates selected CTs against the defined evaluation criteria and proposes the most appropriate for the scope of personal health assistance for elderly people. Chapter 7 summarizes the results and gives an outlook on further work.

2 RESEARCH METHODS

To obtain an overview of the state of the art in CT, a literature review was conducted. We used the keywords “language assistant”, “voice assistant”, “voice interface”, “cognitive assistants”, “intelligent personal assistants”, and “conversational user interface” combined with “comparison” and “analysis” to search for papers in Google Semantic Scholar, Springer Link, IEEE Xplore and ScienceDirect. In addition, we searched relevant conference series and journals for articles that fall within the scope described above. Furthermore, we used references of the retrieved articles to find further works on the topic. The research also had to be extended to the official websites and tutorials of relevant CTs. This resulted in a set of around 50 sources.

The next step was to pre-select, group, install and test the identified CTs with basic use cases. The goal was to get a basic understanding how to use these CTs.

As domain experts, we performed a qualitative analysis and selected relevant categories and criteria for our domain. The CT candidates were then evaluated qualitatively.

3 STATE OF THE ART

Rapidly evolving technologies make CTs a constantly changing field. Recent industry studies conducted by Sumo Heavy Industries (2019), Olson & Kemery (2019), and Siddike et al. (2018) provide an insight of commercial CT representatives. Voice commerce is becoming increasingly important for companies. The market leaders in this field are Amazon’s Alexa and Google Assistant with the highest amount of market share (Sumo Heavy Industries, 2019). The most frequently mentioned commercial representatives are Amazon’s Alexa, Google Assistant, Apple’s Siri, Microsoft’s Cortana and Samsung’s Bixby. In addition to these commercial providers, there are open source providers on the market like Mycroft, Snips and Rhasspy. In contrast to commercial providers, these open source solutions can be analyzed more thoroughly (Kinsella & Mutchle, 2019).

In addition, there are open-source text-to-speech and speech recognition libraries such as CMUSphinx, MaryTTS, and Kaldi that can be used by developers to create CTs from scratch (Gaida et al., 2014). A disadvantage of these libraries is that they do not offer any development environment nor predefined functionalities. Therefore, although a comparison with integrated CTs is incomplete with respect to some criteria, we include them in our analysis.

Only few research papers are focusing the developer’s point of view on a CT. Studies conducted by SUMO Heavy Industries (2019), Siddike et al. (2018) or Olson & Kemery (2019) pay special attention to the end-user’s perspective. But from the developer’s point of view, many factors of a CT can influence the selection, implementation and deployment process.

In the context of our literary research we could not find any work which focuses on the application domain and the end-user needs and follows a criteria-based approach to select an appropriate CT. Besides security and privacy also aspects like the required end user language, trust in the provider or costs are relevant. This paper contributes to fill this gap and shows criteria which are applicable and evaluable for the application domain of AAL support for the elderly.

4 ANALYSIS CRITERIA

In order to find suitable criteria, the criteria from other studies (e.g. Bedford-Strohm (2017), Candia (2017), Kinsella & Mutchle (2019), Olson & Kemery (2019), Siddike et al. (2018) and SUMO Heavy Industries (2019)), and the criteria for health applications (Fraunhofer-Gesellschaft, 2020) were gathered

and combined to remove duplicates. These criteria were then clustered into three main categories: usability for the users, alignment with the needs of the target group and implemented or available security and privacy mechanisms. After briefly introducing the category, each criterion is described in the following in more detail. The resulting list of criteria is not exhaustive, but serves as a tool for selecting CTs for the problem domain.

4.1 Usability

Criteria that focus on the usability of the complete application life cycle are necessary to measure how convenient the interactions between the user and the application are. Without a high usability, any interaction with it will turn out to be tedious and not optimal for any user. This category focuses on the convenience and comfort that is given by the different voice assistants. As such all the criteria, except for “Intelligence” (U6), were deduced with the help of existing literature (Capgemini, 2018; Claessen et al., 2017; Kinsella & Mutchle, 2019; SUMO Heavy Industries, 2019; Wolters, 2015). These, to some extent, already implement such categories, but never consider all of them at the same time.

U1 – Ease of Installation: This criterion checks the effort of commissioning a voice assistant by analyzing both the installation as well as then configuration of the device. The easier the process is designed, the better the voice assistant is rated compared to the others.

U2 – Extensibility: After the initial configuration and installation have been finished, users may want to extend the existing functionality (e.g. stores) in an intuitive and easy way. When comparing the CTs, the intended ways of adding functionality are analyzed and ranked according to their difficulty.

U3 – Comfort of Dialogues: The biggest benefit of a conversational user interface is that they enable interaction through voice, whereas “classical” CT is based on text, which has to be typed. For the voice interaction to be as natural and intuitive as a normal human-to-human conversation, certain characteristics have to be achieved.

U4 – Dynamic Interaction: Without flexibility in the process of developing an application, voice assistants cannot be as dynamic as they need to be. An example for such a restriction is that only the user – and not the voice assistant – is able to start a conversation. The more difficult it is to achieve the desired outcomes, the less likely a voice assistant will turn out to be the most usable from the user’s point of view.

U5 – Accuracy of Queries: When users interact with the conversational user interface, accuracy is necessary to enable the best experience. Especially, in a field like eHealth, if spoken information is misinterpreted, this can cause severe problems that eventually threaten life.

U6 – Intelligence: Intelligence in this context is understood as the ability of a conversational user interface is to react correctly to a wide range of situations. In this case, intelligence was measured in a way, where assistants were asked a set of questions. The results are evaluated with respect to the correct understanding of the sentence and if a suitable answer was provided.

U7 – Multimodality: With this criterion the capability of a voice assistant to handle more than one way of interaction and representation is evaluated. Graphical user interfaces (GUIs) will not be completely replaced by CUIs in the foreseeable future. On the contrary, user tests (Këpuska & Bohouta, 2018) made it obvious that many things can be controlled much more easily with the help of a GUI than only with speech input and output. Multiple forms of interaction (e.g. by means of a microphone, a camera or a display) can significantly increase the usability of a voice assistant, specifically in the case of entering larger amounts of data or by providing additional visual aid (Këpuska & Bohouta, 2018) or for applying different authentication methods. Therefore, the score increases with the number of ways of interaction.

4.2 Target Group: The Elderlies

Criteria for this category were derived from a persona that was specific to the mentioned target group. The underlying observations are described here shortly. Based on the demographic changes, the amount of elderly people in the population increases constantly. Additionally, the amount of age-related diseases rises and therefore also the number of caretakers which are needed. With many positions in this field being unoccupied or underpaid, this could create a situation in which a shortage of caretakers becomes inevitable. Patients want to stay autonomous and want to stay in their own homes as long as possible. Providing assistive technology can help to fulfill this desire and even more can enable the patients to take actively part in the care-taking process. With the help of unobtrusively integrated voice assistants, it is possible to give the elderlies more control and a way to naturally participate in eHealth related fields. With this information it was possible to select the following criteria:

T1 – Costs: The aspect of costs must be covered. The target group is defined in a way that they strive

for a reasonable price for the assistant. To ensure, that all costs are compared on the same basis, they will be analyzed over a timespan of a year, as reoccurring costs need to be considered as well.

T2 – Language Support: Speaking and understanding the language most commonly spoken in the target group is very important concerning the acceptance of voice assistant technology. In the best case, this allows the target group an unproblematic interaction with no misunderstandings because of language barriers.

T3 – The Companies’ Trustworthiness: The trust, users have in the company deploying the CUI has found a lot of attention (Kinsella & Mutchle, 2019). Without trust, the users are reluctant to purchasing or even using a specific voice assistant. As it is difficult to measure how trustful a company is in the eyes of a user of the target group, this criterion focuses on the active contributions of the companies to reduce trust issues.

T4 – Smart Home Integration: Although, CUIs come with a wide array of functionality, the implementation itself can be limiting. As such it is of interest, if the voice assistant technology is capable of being integrated in already existing smart homes, or if it is possible to build a smart home with the assistant as a base. This benefits the target group, as other smart devices can be integrated as well.

T5 – Number of Applications: The number of applications reflects the range of different uses the voice assistant can have. The term “applications” in this context means software that can be installed beforehand or after setting up the voice assistant. For analysis, the number of predefined applications in the corresponding app-market is counted. Besides the number of applications, the variety of these is the second factor positively influencing the score.

T6 – Support of Other Devices: This criterion covers all the possible ways for deploying the CUI. Every form of end-device is considered, such as smartwatches, smart TVs, smartphones and proprietary devices. The greater the diversity, the better the assistant is rated in terms of this criterion.

4.3 Security and Privacy

As applications in the area of eHealth and AAL process personal, and especially medical data (i.e. critical data with respect to GDPR, (European Parliament, 2016, Article 9), special attention must be paid to the security and the privacy of this data. Besides laws and regulations, security and privacy of medical information have a high priority, as users do not want their data to be leaked, misused or even sold (Kinsella &

Mutchle, 2019). In order to address these concerns, the following criteria focus on some of the main issues gathered throughout the literature research (e.g., Alepis & Patsakis (2017), Candid (2017), Gong & Poellabauer (2018), Kang et al. (2019), Kinsella & Mutchle (2019), Olson & Kemery (2019) and Tiropanis et al. (2015)). Again, these criteria are to some extent already observed by these papers and are reoccurring across them. Nonetheless, never are all categories considered at the same time.

S1 – Domain Specific Infrastructure: To ensure the highest control over the processed data, the use of personal infrastructure is recommended. This adds a layer of security, as permissions, authentication, and hardware can be controlled by the user itself, or a trusted admin. The necessary interfaces must be provided by the voice assistants, as well as no restrictions by the system must be mitigating the overall control of the infrastructure. Specifically, the location of the database is checked and evaluated.

S2 - Location of Voice Commands: When a user issues a command, this command may be stored in the process. Possible locations for the stored data may be locally on the device, on a server or in a (vendor-specific) cloud. As it is personal/medical data, the preferred outcome would be that information is not stored permanently. Ideally, the temporary data is deleted, after the request has been processed.

S3 - Location of Intelligence: Obviously, natural language queries issued by the user must be digitized to be processed. To achieve this, a voice assistant can either directly process the query on the device or transmit the voice data to an external infrastructure, where an additional way for attacks is created.

S4 – Security of Data Transmission: Ideally, no data is sent to remote services, and all data is processed in the local infrastructure or the CUI itself. But if data is sent over the (untrusted) internet to remote devices (e.g. cloud-based infrastructure), it must be protected in terms of confidentiality, integrity and authenticity. Otherwise, the data can be exfiltrated, manipulated or forged by attackers.

S5 – Access to Stored Information: This criterion focuses on the data intruders would have access to, if they are able to physically seize the assistant. Forms of extracting data would be to use the interfaces that allow direct access to stored data or by speaking to the assistant and receiving information that way.

S6 – Authentication: As not every person should have complete access to all a voice assistant’s functionality, this criterion checks the ways of authentication, a voice assistant offers. Typically, passwords or

PINs are used, but it is also possible for a voice assistant to distinguish between different voices.

S7 – Compliance with GDPR: With the GDPR taking effect in 2018, it is necessary to consider if the voice assistants align with the current law.

Although not every aspect of a technology can be covered with these criteria, a basis for domain-specific evaluation has been established with these criteria. The extension by more criteria could increase the benefits even further. In addition, it may be necessary to weight different criteria for specific use cases, which is explained in more detail in Chapter 6.

5 CT IN THE EHEALTH DOMAIN

Elderlies typically have moderate digital literacy, are often not very financially resilient, sensitive to their personal data and do not want to constantly change the used technology.

Before analyzing the CTs mentioned in chapter 3 to the criteria stated in chapter 4, we checked them against the hard criteria “costs” and “sustainability on the market”. As not every technology fits into the scope of this paper, this selection was necessary. In addition, not enough information could be obtained on all the technologies mentioned, so that a representative set of CTs had to be found for an analysis:

Amazon’s Alexa and **Google Assistant** prove to have the highest usage rate. They therefore had to be considered as they are sustainably represented in the market, affordable and quite well documented.

Apple’s Siri and **Microsoft Cortana** make up a high volume of devices compared to other providers. But Cortana is not further analyzed in depth as Microsoft announced in early 2019, that they will not compete with other voice assistant providers any longer (Novet, 2019) and have no intention to develop their voice assistant in the future. Siri on the other hand has a different drawback. Apple is still competing with other CT providers but restricts the devices Siri and her extensible functionality can be developed and deployed on (Apple, 2020b). With a modular and open system in mind, the users cannot be constrained to use a singular form of device type. Apple devices also tend to be relatively expensive, whereas Google and Alexa both offer cheaper solutions (Amazon, 2020a; Apple, 2020a; Apple, 2020b; Google, 2020b).

Samsung’s Bixby only runs on Samsung devices and occupies only a very small market share. This limits the applicable end devices too much and therefore Bixby is not further investigated in this study.

Can open source CTs compete with commercial providers? To answer this question Mycroft and

Rhasspy were added to the study, though they occupy only a small market share. **Mycroft** was chosen, as it is one of the only open-source solutions, which comes with an already build-in hardware. The special feature of Mycroft is that all the natural language components are modular and allow the developer to change them independently (Mycroft, 2020b).

Rhasspy on the other hand offers no prebuilt hardware and only comes in the form of a set of services which can be freely changed. The advantage of Rhasspy is that it is open source software that allows the developer to create an assistant that runs entirely locally on an end device. This fact makes it interesting for use in eHealth and thus included in the comparison. Note that Snips was previously analyzed in the work of Jesse (2019) but is not freely available anymore and not considered anymore.

In addition, to these complete technology packages for the development of CUIs, also the speech-recognition toolkit **CMUSphinx** and the text-to-speech library **MaryTTS** have to be mentioned. Both offer a lot of freedom, but they are not integrated development technologies and must be configured and built into a solution by a developer. MaryTTS itself is not recommended to be run on low-end computers (e.g. Raspberry Pi) and therefore must be installed separately on a server (CMUSphinx, 2020a; DFKI, 2020).

It has turned out that Kaldi, CMUSphinx and MaryTTS are used as a component of Rhasspy. Hence Kaldi was not pursued further, and the focus was laid on CMUSphinx and MaryTTS as to minimize redundancies in the study (Coucke et al., 2018).

In order to be able to compare CMUSphinx and MaryTTS in combination with the complete technology solutions, a demo assistant was developed to simulate how the two libraries would perform. The specification of the demo voice assistant was taken from Jesse (2019) and used to evaluate the criteria.

6 EVALUATION

This chapter describes the evaluation results of the selected voice assistants based on the set of criteria proposed in Chapter 4. Important results are described in more detail in order to provide the necessary information why a certain result was achieved. More general findings are only discussed briefly but depicted in the Tables 1, 2 and 3. The evaluation is based on the work of Jesse (2019) and put into a new perspective. Note, that for some criteria an implementation of CMUSphinx and MaryTTS was used (combination). This is because without the fictional application no

evaluation of CMUSphinx and MaryTTS could have been conducted. Overall, the same basis is used, and a smart speaker is compared where necessary.

A scoring system was applied, which ranked the different voice assistants on a scale from 1 (optimal solution) to 5 (being the worst out of the observed technologies). Criteria, where no ranking was needed, were evaluated on the criterion being meet (✓) or not (✗). If no suitable information was obtainable or could be found a dash (-) was used.

6.1 Results on Usability

Table 1: Results on usability.

Usability	U1	U2	U3	U4	U5	U6	U7
Alexa	✓	✓	1	2	2	2	3
Google Assistant	✓	✓	1	2	2	1	2
Mycroft	✓	✓	3	2	-	-	1
Rhasspy	✗	✗	2	1	1	-	1
Combination	-	✗	-	1	-	-	1

U1 – Ease of Installation: Alexa, Google Assistant, and Mycroft offer a more user-oriented approach in comparison to the other assistants. Rhasspy has a quite complicated installation process, which has to be performed by a technically skilled person. Concerning CMUSphinx and MaryTTS, no reasonable evaluation was possible because the libraries do not define the process of installation.

U2 – Extensibility: As seen before, Alexa, Google Assistant and Mycroft try to be as user-centered as possible and therefore allow them to add functionality. On the other hand, Rhasspy and CMUSphinx and MaryTTS do not allow for any functionality to be added after the initial setup.

U3 – Comfort of Dialogue: Alexa and Google Assistant offer a high level of comfort during conversations. Rhasspy offers a similar functionality but has no follow-up mode, so that the wake-word must be repeated for every step in a dialog. Mycroft misses a multitude of these features. Concerning the combination, no reasonable evaluation was possible, as it is defined through the implementation of the developer.

U4 – Dynamic Interaction: As Klade (2019) showed, all commercial CTs come with the same amount of restrictions. They do not allow an out-of-the-box initiation of a conversation by the assistant itself. The user must trigger each conversations and reminder become cumbersome. Rhasspy and the two libraries have an advantage, as they do allow for the application to be designed by the developer. Therefore, it is possible to initiate a conversation by those.

U5 – Accuracy of Queries Understood: Finding a comprehensive study that covers all selected CTs

was not possible, therefore the one covering the most assistants were selected. The study of Coucke et al. (2018) covered the NLP components offered by Snips, Google, Alexa, Facebook and Microsoft. The results point towards Snips being the most accurate, which is also implemented in Rhasspy.

U6 – Intelligence: The study conducted by Loupventure (2018) ranked the Google Assistant as the CTs with the highest intelligence. They tested the assistants on the correct interpretation of a phrase and on providing suitable answers. A study testing all observed CUIs could not be found.

U7 – Multimodality: All mentioned CTs allow for multimodality. Open source approaches offer a higher number of possible interfaces than proprietary ones. Actually, open source solutions can be used to create any interaction the developer can imagine, provided the developer is able implement it. If the natural language components of the Google Assistant are used, they also enable a multitude of possible interactions. Alexa on the other hand is more restrictive in the offered ways of interaction.

6.2 Results on the Target Group

Table 2: Results on target group.

Target group	T1	T2	T3	T4	T5	T6
Alexa	1	✓	3	1	1	1
Google Assistant	2	✓	3	1	2	1
Mycroft	4	✗	2	2	3	3
Rhasspy	3	✓	1	2	4	2
Combination	5	✓	1	2	-	2

T1 – Costs: The commercially available low-end smart speakers are cheaper than any open-source smart speaker solution. Rhasspy and the two libraries require a Raspberry Pi and additional hardware. This hardware cumulates a significant amount of money and they are therefore ranked lower.

T2 – Language Support: In order to fulfill this criterion in this study, German must be supported. Studies using this analysis may adjust the selected language according to their needs. All conversational interfaces, except for Mycroft, work with German. Mycroft does not officially support German.

T3 – The Companies’ Trustworthiness: In comparison to Alexa and Google Assistant, which collect information on the user to improve their services, Mycroft, Rhasspy, and CMUSphinx and MaryTTS do not per default. However, Mycroft offers an opt-in policy that allows users to decide, whether their data can be used to train the technology.

T4 – Smart Home Integration: When a CT is deployed on an end device, it is possible to combine

and connect other technology (applications) to it. Mycroft, Rhasspy, and CUI libraries enable this tendentially better, as they are developed with an open-source approach in mind. There can be changes made to the main application and the interfaces (CMUSphinx, 2020a; Mycroft, 2020b, Rhasspy, 2020). This enables the creation of CTs that are much more flexible than proprietary solutions.

T5 – Number of Applications: The better known a technology is, the more applications are in its markets. Hence, the market leaders offer a higher variety and number of applications to be installed than other open-source or less known solutions.

T6 – Support of Other Devices: In terms of sheer amount of deployable end devices, Alexa and Google Assistant offer the biggest variety of already finished solutions. As they sell a variety of different devices, they outperform the other CUIs in that regard. These other CUIs (Mycroft, Rhasspy, etc.) mainly focus on providing the means for developers to create a CT.

6.4 Results on Security and Privacy

Table 3: Results on security and privacy.

Security	S1	S2	S3	S4	S5	S6	S7
Alexa	✓	4	3	2	1	1	✓
Google Assistant	✓	3	4	2	1	1	✓
Mycroft	✓	2	2	2	2	2	✓
Rhasspy	✓	1	1	1	2	2	✓
Combination	✓	2	2	1	2	2	✓

S1 – Domain Specific Infrastructure: All observed CUIs allow the integration of infrastructure. Open-source solutions pose an inherently more difficult process than the other ones, as everything must be implemented by the developer.

S2 – Location of Voice Commands: Rhasspy is the only solution that can fully process the voice commands on the device and never transmits or stores requests. Mycroft and CMUSphinx and MaryTTS do not store voice commands explicitly, but do transfer them to a remote system for processing. Alexa and Google Assistant use the received voice commands and store them for later use and training.

S3 – Location of Intelligence: As seen before, Rhasspy can process everything locally, which makes it the best choice for this criterion. Mycroft needs to send data to a remote system for certain NLP components to work. Alexa provides basic functionality without using remote services, but otherwise sends all voice data to a cloud. Google Assistant needs to send all information to a server because it cannot process any of them on the device.

S4 – Security of Data Transmission: Like the two previous criteria, Rhasspy does not require any security measures for data transmission because no information is transmitted. The other CTs encrypt and secure their data transfers appropriately (SSL/TLS), but still do not have the ability to work fully on the device to remove a potential point of attack. Therefore, they must be ranked lower than Rhasspy for this criterion.

S5 – Access to Stored Information: Alexa and Google Assistant mainly use proprietary devices with no accessible local storage. Any application associated with such a commercial solution must be authorized by the user. Open source solutions have the disadvantage that they have a base like a Raspberry Pi.

S6 – Authentication: Concerning Authentication, Alexa and Google Assistant stand out, because they both have a built-in way of authentication but can be extended with others (Auth0, 2020). Hence, they are rated higher. The open-source technologies do not offer any built-in authentication, but it can be implemented by developers.

S7 – Compliance with GDPR: All CTs claim to align with GDPR and other laws concerning the protection of personal data. It can be emphasized that Rhasspy mentions its offline functionality directly on their landing page, whereas other technologies do not mention if they align to the GDPR. Therefore, it was necessary to analyze their security statements and policies, in which this information could be found.

7 SUMMARY AND CONCLUSION

In this article, we have defined evaluation criteria to select the most appropriate dialogue-oriented AI technology for the area of personal health management of the elderly. Based on these criteria, we evaluated and ranked the leading commercial and open source CTs for this application domain.

At a first glance, Alexa and Google Assistant perform very well in this evaluation. But in terms of sensitive health data, security and data protection in the application domain under consideration, they are not an alternative for our problem domain. Overall, Rhasspy showed to be superior in regard to the criteria of “Accuracy of understood queries” (U5), “The companies’ trustworthiness” (T3), “Location of voice commands” (S2), “Location of intelligence” (S3) and “Security of transmission” (S4). This suggests that Rhasspy is the best choice, especially in terms of security and data protection. Since the technology is able to work completely “out of the cloud”, no other analyzed technology can outperform Rhasspy in our

eHealth domain. This makes Rhasspy the best CT choice to create a voice assistant focused on working with critical health data.

Nevertheless, criteria like “Ease of installation” (U1), “Extensibility” (U2) and “Number of applications” (T5) show the downside of Rhasspy. Here, it is clearly outperformed by Alexa and Google Assistant. However, since the functionality and the functioning of these technologies are constantly changing, it is necessary to carry out regular re-evaluations.

The research presented in this paper was carried out as part of the AYUDO project (FFG Projektdatenbank, 2020), which is intended to help the elderly improving or preserving their health using a CUI. Beyond that project the criteria can also form a basis for the evaluation of CTs to be used in other application domains. Here, some specific criteria will have to be added to the existing categories, and/or the weight of single criteria has to be adapted to the requirements of the application domain.

REFERENCES

- Alepis, E., Patsakis, C. 2017. Monkey Says, Monkey Does: Security and Privacy on Voice Assistants. In *IEEE Access* 5, pp. 17841–17851. DOI: 10.1109/ACCESS.2017.2747626.
- Amazon.com, Inc. (2020a, Nov. 20). Built-in Devices. www.amazon.com/b?node=15443147011.
- Amazon.com, Inc. (2020b, 11/2020). Create Skills for Alexa-Enabled Devices with a Screen. <https://developer.amazon.com/de/docs/custom-skills/create-skills-for-alexa-enabled-devices-with-a-screen.html>.
- Apple Inc. (2020a, Nov. 20). Apple HomePod. www.apple.com/de/shop/buy-homepod/homepod.
- Apple Inc. (2020b, Nov. 20). Apple Homepage. www.apple.com.
- Auth0. (2020, Nov. 20). Auth0 Overview. <https://auth0.com/docs/getting-started/overview>.
- Candid W., (2017). A guide to the security of voice-activated smart speakers. Symantec Corporation.
- Capgemini. (2018). Conversational Commerce. Why Consumers Are Embracing Voice Assistants in Their Lives.
- Claessen, V., Schmidt, A., Heck, T. 2017. Virtual Assistants. In: Humboldt-Universität zu Berlin.
- CMUSphinx (2020a, Nov. 20). Tutorial For Developers. <https://cmusphinx.github.io/wiki/tutorial/>.
- Coucke, A., Saade, A., Ball, A., Bluche, T., Caulier, A., Leroy, D. et al., (2018). Snips Voice Platform: an embedded Spoken Language Understanding system for private-by-design voice interfaces.
- DFKI GmbH. (2020, Nov. 20). Mary Text To Speech <http://mary.dfgi.de/documentation/overview.html>.
- European Parliament 2016. General Data Protection Regulation. <http://eur-lex.europa.eu/eli/reg/2016/679/oj>, pp. 1–88.
- FFG Projektdatenbank. (2020, Nov. 20). AYUDO. <https://projekte.ffg.at/projekt/3311832>.
- Fraunhofer-Gesellschaft. (2020, Nov. 20). Appkri – Kriterien für Gesundheits-Apps. <https://chealth-services.fokus.fraunhofer.de/BMG-APPS/categories/Alle>.
- Gaida, C., Lange, P., Petrick, R., Proba, P., Malatawy, A., Suendermann-Oeft, D., (2014). Comparing open-source speech recognition toolkits. In: Technical Report of the Project OASIS.
- Gong, Y., Poellabauer, C. 2018. An Overview of Vulnerabilities of Voice Controlled Systems.
- Google LLC. (2020a, Nov. 20). Development. <https://developers.google.com/actions/overview>.
- Google LLC. (2020b, Nov. 20). Google Store. <https://store.google.com>.
- Hoy, M., (2018). Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants. In *Medical Reference Services Quarterly* 37. DOI: 10.1080/02763869.2018.1404391.
- Huang, X., Acero, A., Hon, H., (2001). *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. 1st. Upper Saddle River, NJ, USA: Prentice Hall PTR.
- Jesse, M. W. 2019. Analysis of voice assistants in eHealth. Master Thesis. DOI: 10.13140/RG.2.2.14691.37924.
- Kang, B. B., Jang, J., Cho, G., Choi, J., Kim, H., Hyun, S., Ryoo, J. 2019. Threat Modeling and Analysis of Voice Assistant Applications. In *Information Security Applications*. Cham: Springer International Publishing.
- Képuska, V., Bohouta, G., (2018). Next-generation of virtual personal assistants. 2018 IEEE 8th Annual Computing and Communication Workshop and Conference.
- Kinsella, B., Mutchle, A., (2019). Smart Speaker Consumer Adoption Report.
- Klade, J. 2019. Sprachassistenz zur Patientenunterstützung. Master thesis.
- Loupventure. (2018). Annual Smart Speaker IQ Test.
- Mycroft AI, Inc. (2020a, Nov. 20). Privacy Policy. <https://mycroft.ai/embed-privacy-policy/>.
- Mycroft AI, Inc. (2020b, Nov. 22). Software and Hardware. mycroft-ai.gitbook.io/docs/mycroft-technologies.
- Novet, J. (2020, Nov. 20). Newsreport www.cnn.com/2019/02/05/google-no-longer-considers-microsoft-cortana-a-competitor.html.
- Olson, C., Kemery, K., (2019). Voice report. From answers to action: customer adoption of voice technology and digital assistants. Microsoft, Bing. <https://about.ads.microsoft.com/en-us/insights/2019-voice-report>.
- Rhasspy. (2020, Nov. 24). Documentation. <https://rhasspy.readthedocs.io/en/v2.4.20/>.
- Siddike, Md. A., Spohrer, J., Demirkan, H., Kohda, Y., (2018). People’s Interactions with Cognitive Assistants for Enhanced Performances. SKIM: Voice Tech Trends 2018.
- SUMO Heavy Industries. (2019). 2019 Voice Commerce Survey.
- Tiropanis, T., Vakali, A., Sartori, L., Burnap, P., McMillan, D., Loriette, A. 2015. Living with Listening Services:

Privacy and Control in IoT. Internet Science. Cham: Springer International Publishing.

Wolters, M. K., Kelly, F., Kilgour, J., (2015). Designing a spoken dialogue interface to an intelligent cognitive assistant for people with dementia. In Health Informatics Journal 22 (4), pp. 854–866.

