Generic concept for integrating voice assistance into smart therapeutic interventions

Jens Scheible[1], Fabian Hofmann[1], Manfred Reichert[1], Rüdiger Pryss[2], Marc Schickler[1]
[1]*Institute of Databases and Information Systems, Ulm University, Germany*
[2]*Institute of Clinical Epidemiology and Biometry, University of Würzburg, Germany*
[1]{*jens.scheible, fabian-1.hofmann, marc.schickler, manfred.reichert*}*@uni-ulm.de*
[2]*ruediger.pryss@uni-wuerzburg.de*

***Abstract*—*Therapeutic Interventions (TIs)* play an important role in modern medical and psychological treatments, but their integration into the digital world still shows deficits, e.g., in the integration of the auditory interface. Initiatives to integrate this interface into existing *Internet- and Mobile-Based Interventions (IMIs)* are largely focused on a small group of *Voice Assistants (VAs)* and their specific capabilities. To mitigate these drawbacks, the presented concept seamlessly integrates arbitrary VAs into the treatment process of TIs. To this end, an architecture - including a discussion of relevant requirements - is presented that, on the one hand, uses VAs as the only point of contact with patients and, on the other hand, provides a comprehensive web-based backend for *Healthcare Providers (HCPs)*. Based on the architecture, a proof-of-concept implementation using *Amazon Alexa* is presented. Finally, it is discussed that the scenario addressed and the solution presented have great potential, but still need a lot of work and technical considerations.**

***Index Terms*—conversational agents, smart assistance, therapeutic interventions, voice interface**

I. Introduction

NOWADAYS, *Therapeutic Interventions (TIs)* play an essential role in medical and psychological treatment. Their scope of application ranges from medication administration to complex therapeutic homework [1]. The demand for digital solutions has therefore risen sharply in recent years and has led to numerous, often multimodal concepts and developments [2], [3]. However, these developments have mainly focused on visual user interfaces provided via smartphones, tablets or wearables. With the rapid development of *Voice Assistants (VAs)* capabilities, their importance has now increased, creating a completely new way of interaction with patients.

In general, it does not seem trivial to reconcile the two worlds of VAs and TIs, especially considering the complexity of both fields [4]. To be more precise, TIs consist of a wide range of different methods and tasks that are tailored to the individual needs of individual patients. VAs, on the other hand, are offered by different companies such as Amazon or Google; in addition, many research approaches exist, for which the available solutions are characterized by a large heterogeneity. Consequently, the difficulty in bringing them together lies not in digitizing TIs, but in creating a uniform and comparable interface for any VA for TIs. In order to make auditory interfaces available through VAs in the context of TIs, we believe that further research is needed. Above all, one aspect is particularly important: a developed concept for the integration of VAs into TIs must not lose its genericity with respect to a provider of VAs. Furthermore, possible concepts should not only focus on patients, but also integrate therapists, domain specialists and scientists in a suitable way.

To enable this, from a technical point of view, the digital delivery of interventions is considered as the starting point for our approach. Therefore, VAs are considered as agents that only access treatment data. In addition, typical methods of language dialogue systems, such as *Natural Language Understanding (NLU)* and *Natural Language Generation (NLG)* [5], are used to leverage their generated information to better enable more personalized user interactions. In this process, agents are orchestrated by a server. The contribution of this work is to present a *(1)* generic concept for the integration of VAs in therapeutic contexts. Within this concept, a corresponding architecture is presented. Furthermore, the *(2)* challenges and limitations of such an architecture are discussed and a *(3)* prototypical implementation based on this architecture is presented.

The remainder of this paper is organized as follows: Section II discusses related work. Key aspects of TIs are presented in Section III, while requirements for the proposed approach are presented in Section IV. The development of the approach and its implementation are discussed in Sections V and VI. A discussion of the results, benefits, and limitations is presented in Section VII. Finally, Section VIII concludes the paper with a summary and outlook.

II. Related Work

Previous approaches integrating a voice interface into *Therapeutic Interventions (TIs)*, mostly concentrate on single *Voice Assistants (VAs)*. According to an up-to-date (2020) market share[1], *Amazon Alexa* [6] (34%) and *Google Assistant* [7] (43%), still occupy a prominent position in the world of VAs. Due to the ability to bundle a number of voice commands into so-called *skills* that have, for example, a common dialog context, as well as the ability to easily integrate other systems easily, *Alexa* is very popular in the present context. Of note, the developer API of *Alexa Skills Kit*[3] was launched in 2015, while for *Google Assistant*, *Google Actions*, in 2017[2]. Of further note, existing platforms for *Internet- and Mobile-Based Interventions (IMIs)* [8]–[10], which are important technical foundations for TIs, support patients as well as *Health Care*

Providers (HCPs) during the overall intervention progress. The resulting setting must therefore be carefully considered in the present context if VAs are to be used for TIs. The following sections first introduce a number of platforms that implement therapeutic interventions in the eHealth context and briefly highlight their capabilities. Afterwards, identified approaches that integrate the auditory interface into the digital interventions will be presented.

A. Therapeutic Intervention Platforms

The eSano platform [8], as an example of IMIs, offers treatment for mental disorders and chronic somatic diseases independent of time and place. The platform consists of a central REST API that connects a database with web-based systems to flexibly create interventions. Patients can perform their treatments via various cross-platform applications. The applications support diaries, questionnaires, and *Ecological Momentary Assessments (EMAs)*, among others. The authors of [9], in turn, provide a messaging system that patients can use to communicate with their therapists. For example, patients can book appointments or provide feedback on ongoing treatments. In addition to this asynchronous type of communication, real-time chat functionality has been added to enable live communication via text, voice or video. Because of the features available to patients and HCPs, both platforms provide a solid starting point for integrating the auditory interface. However, it needs to be clarified to what extent the platform APIs are sufficient as interfaces that VAs are able to interact with.

Electronic Health Records (EHRs), stored in a centralized medical cloud, are a key feature of [10]. The authors describe how the collection of physiological data can be integrated into their system, e.g., by collecting electrocardiological data via wearables. A connected smartphone provides synchronization and allows healthcare professionals to draw conclusions about, for example, daily activity via a web interface, which supports clinical decision making.

B. Enabling Interventions for VAs

Yun et al. [2] evaluated the integration of voice interfaces in *Caring for Caregivers Online (COCO)*[3]. COCO is a platform focusing on private caregivers of family members with chronic health conditions. The majority (89%) of participants in the survey conducted are familiar with VAs, and the fact that a voice interface provides more flexibility in retrieving information or instructions for practices during activity favors VA use. The authors also cite high levels of user satisfaction, such as not being distracted by a smartphone while interacting with the VA.

[2] evaluate the degree of *User Experience (UX)* with the help of user testing methods and could reveal challenges in the following three categories: (1) *Accessibility:* Since users prefer learning a new system in combination with a *Graphical User Interfaces (GUIs)*, keeping track of the system progress and controlling is rather hard only using spoken language interfaces. (2) *Efficiency:* Voice as a volatile form of information should be accompanied by a reduced cognitive load. Presented information should not be too detailed, nor should it lose too much information content. (3) *Satisfaction:* Users prioritize a high level of correctness over speed of response in this context.

Similar to [2], the authors of [11] integrated a voice interface into an existing IMI platform: *Nurse AMIE* [12]. The system supports women with metastatic breast cancer through a self-administered intervention solution that includes psychoeducation and mindfulness meditations. It technically relies on *Alexa* and the corresponding Amazon infrastructure (i.e., AWS Lambda Service, AWS DynamoDB). Previous approaches provide proven interventions using conventional web- or mobile based interfaces. Underlying infrastructures start from those of single manufacturers, potentially limiting the use of different VAs [11], to completely decoupled distributed systems [8]–[10]. As the auditory interface is on the rise[4] and gives new possibilities in a more natural human computer interaction, we think this field of research is promising. Different approaches already try to tackle challenges like e.g., integrating distributed systems into multiple VAs [10].

However, the aforementioned approaches lack their generic character, as they focus on specific VAs or user groups. The concept presented here, on the other hand, attempts to include all stakeholders in order to develop a more general approach to integrating VAs into the TI domain. VAs are therefore considered as interchangeable voice interfaces.

III. Therapeutic Interventions

The starting point of any therapeutic intervention is the occurrence of a psychological or physical complaint, which leads to consultation with a corresponding specialist. In the further course, a diagnosis can be made using various methods. This diagnosis can in turn serve as the basis for planning any necessary treatments [1]. Traditional therapeutic interventions are based on one or more therapy sessions [1]. These serve to refine the initial diagnosis and to test possible therapeutic problem-solving approaches together with the patient. The treatment itself consists mostly of both inpatient and outpatient measures [13]. During a supervised session, for example, the patient performs exercises under the guidance of a therapist. However, to achieve optimal therapy results, exercises must be continued outside of these sessions. Exercises that the patient performs independently outside of the sessions are referred to as *therapeutic homework* [13] and aim to further deepen his or her knowledge and understanding of the therapy.

[1]https://www.statista.com/statistics/789633/worldwide-digital-assistant-market-share/ (Accessed: 2022-04-06)

[2]https://www.theverge.com/2017/5/17/15648538/google-actions-android-ios-phones-third-party-app-assistant-io-2017 (Accessed: 2022-04-06)

[3]https://press.aboutamazon.com/news-releases/news-release-details/amazon-introduces-alexa-skills-kit-free-sdk-developers/ (Accessed: 2022-04-07)

[3]https://coco.health (Accessed: 2022-04-06)

[4]https://www.thinkwithgoogle.com/consumer-insights/consumer-trends/voice-technology-purchasing/ (Accessed: 2022-04-07)

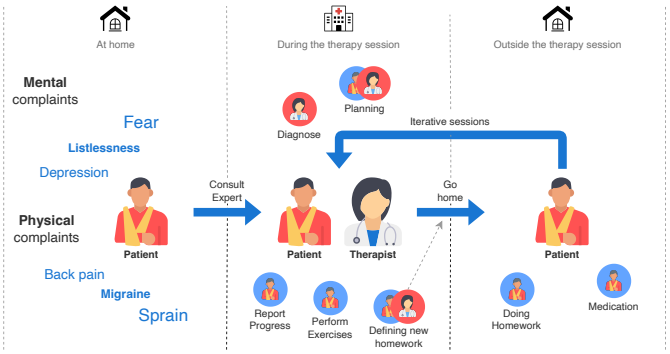

Fig. 1: Traditional therapeutic intervention

However, the success of a therapeutic measure depends not only on the methodology used. Numerous studies show that patient adherence (also compliance) is an essential factor for the success of a therapy. This is because, despite correct indications, problems can occur when performing therapeutic exercises [1]. For example, various challenges may arise in the application of therapeutic homework, such as:

- Misunderstandings in oral communication
- Inappropriate difficulty levels of the tasks
- Lack of motivation

The situation is further complicated by the fact that homework carried out in analog form is difficult to monitor. Digital support of therapeutic processes therefore offers clearly identifiable advantages on both the therapist and patient sides. Therapists could make ad-hoc adjustments in the course of therapy and patients could benefit from even more individualized care.

IV. Requirements

After an initial introduction to *Therapeutic Interventions (TI)*, the following section focuses on the requirements that arise in such a context. When integrating *Voice Assistants (VAs)* into the TI domain, therapists and patients can be considered as the key stakeholders. In addition, to meet the requirements of a generic approach, other stakeholders such as domain specialists or scientists are also considered. Since patients and therapists directly influence therapy outcomes, whereas scientists or domain experts usually influence therapy more indirectly, both perspectives are considered separately. However, the focus of our approach is on the former perspective. Finally, Table I summarizes identified requirements, grouped by the main implementation steps.

Providing the best possible care to a patient is undoubtedly the main goal of any treatment. Ensuring this goal during a therapy session depends on the therapist's skills [3]. Other parts, however, such as homework, take place outside a session and are therefore within the responsibility of the patient [13]. The out-of-session scenario is particularly well suited to the use of a VA, since the VA can serve as a source of information here ①. However, since speech is a volatile form of information [2], it should be supplemented by a visual representation. This, in turn, can be used to balance the level of detail and relevance of individual pieces of information. For example, a VA could provide a visual overview of current progress or answer simple questions auditorily. To increase a patient's adherence, it is also beneficial to gain the user's trust by making the conversations as natural as possible ④. This could be achieved by enriching conversations with historical patient data, current contextual information, or possible predictions ⑤. Thus, the context awareness of a VA agent depends on the available contextual data and the orchestration of the server.

Another important perspective is that of the therapist, who is involved in the planning, implementation and evaluation of a particular treatment. The involvement of a VA at this point therefore means supporting an entire process chain. The coordination of session appointments by a VA is therefore only one essential area of application. In this scenario, an assistant could also monitor the course of treatment, inform the therapist of any kind of deterioration, or simply allow therapists to access current patient data ①. However, it should be kept in mind that just because something can be augmented by a VA does not mean that this is the appropriate modality. For example, consider the visualization of the vast amounts of data generated in the course of therapy. Here, it must be doubted whether a VA is the right modality for the data volumes. Also, the visual support of ad-hoc adjustments of the patient's therapeutic homework may usually be clearer on a large desktop application window. Regardless of the implementation chosen, however, consistency and timeliness of treatment data should also be a key requirement. In addition, different user roles with corresponding authorization levels are already emerging in this scenario ③. This is an aspect that must also be taken into account when integrating a VA into the TI domain.

The last perspective considered is that of researchers and professionals. While therapists and patients are actively involved in a particular treatment, researchers and other professionals tend to take a passive role. Thus, their main interest is initially focused on the data collected [3]. For example, scientists could access the raw data to gain new insights or discover new therapies. This could be supported by making data available in standardized, well-known formats to increase interoperability with external analysis tools ⑦. Professionals could process this data for any type of business transaction. This, in turn, will lead to new, previously unknown requirements in the future. Therefore, the underlying system architecture should be as modular as possible ⑥, and also highly extensible. Eventually, new data-processing or -evaluating components could be extended, VAs could be replaced, or new ones could be added.

V. Architecture

After evaluating some key requirements, the following chapter presents our architectural approach for flexible integration of *Voice Assistants (VAs)* into the therapeutic context.

As mentioned before, the VAs can be considered as agents controlled by a central server. Therefore, the proposed archi-

No.	Title	Description
A.) Interaction Model		
①	Data access	VA gives access to most recent information about treatment.
②	Presentation	According to hardware-limitations, the VA presents answers in an audio-visual manner.
③	User roles	User roles allow a limitation of access for single features.
B.) Conversational Flow		
④	Flow control	Contextual information throughout a dialog influences its outcome.
⑤	Personalization	VA is aware of historic data and current contextual information.
C.) Data and Request Processing		
⑥	Modularity	The modular and extendable system structure benefits further adaptions.
⑦	Interoperability	Internal formats are based on state-of-the-art standards to provide the highest degree of interoperability.

TABLE I: Summary of identified requirements, grouped by A.) Interaction Model, B.) Conversational Flow and C.) Data and Request Processing of *Voice Assistants (VAs)*.

tecture design is based on a well-known client-server model. The main business logic is initially centralized on a single server. Accordingly, the main focus of this approach is on the internal server architecture.

First, the various existing clients must be evaluated in order to know what a server must be capable of to interact with these clients. For example, the server must first and foremost be able to cope with a heterogeneous landscape of different VAs. But, as described earlier, it must also be considered that for some applications the voice interface may not be the appropriate implementation. Therefore, desktop or web applications should also be considered as possible clients. To manage this large number of endpoints, we propose to divide them into voice assistants and other client devices. However, the focus is clearly on the former. The server must therefore provide two separate interfaces to meet the different requirements of the individual device groups.

Due to the variety of VAs currently available, it is not possible to consider all voice assistance implementations. Therefore, according to the latest statistics in [14], only a subset is considered, covering most of the market. Since some of them focus only on the Asian market, such as Baidu, Xiaomi or Alibaba, they are not further considered. In addition, some manufacturers, such as Apple, only offer limited access to the assistant's features, so they are not considered either. Ultimately, the architectural design focuses on Google's Assistant and Amazon's Alexa as reference VA implementations. These two assistants are also particularly suitable due to their large market share in [14], extensive documentation [7] [15], and large developer community.

The first interface to be defined is the VA facing side of the architecture. After it is known what kind of assistants the server has to support, the VA interface can be designed. Our two references VA Google Assistant and Amazon Alexa are

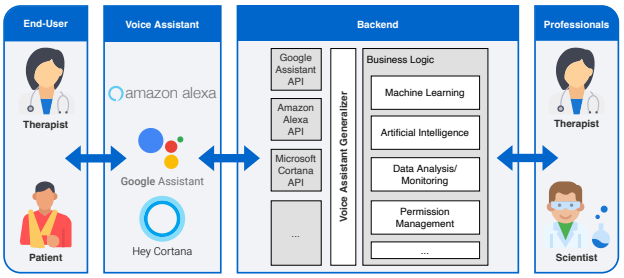

Fig. 2: Schematic description of the architecture

based on the definition of a so-called interaction model [15], [16]. These models describe the main interactions between a user and the assistant. This includes, for example, what a user or patient can say to express certain intentions, which in turn can be analyzed by the conversational agent via *Natural Language Understanding (NLU)*. In the case of reference VAs, this process is provided via an encapsulated, vendor-specific cloud service. Thus, only the result of the NLU, the identified intent of the user, can be accessed. Therefore, the main task of the server is to understand these intentions and provide personalized responses. Generalizing a user's intention represented by VA requests is not a trivial task. Especially when considering the previously mentioned heterogeneous feature set of the individual assistants. Therefore, to ensure that the server is able to handle all incoming requests from the VAs, it is necessary to define possible user interactions in advance. This could be done by modeling a unified interaction model as a basis for structuring the server interface.

To keep the proprietary parts of the architecture as lean as possible, the incoming VA requests need to be translated into an internal, vendor-independent format. This could be done by using an intermediate layer that acts as an intermediary between the internal logic and the incoming VA requests. Therefore, each assistant gets its own communication interface that implements the basic vendor-specific communication logic. The request and its fulfillment is delegated to the mediator. Subsequently, the mediator invokes the actual business logic and generates the corresponding response.

To keep the mediator logic as simple as possible, we propose to use established and well-known data formats that can be interpreted by the assistant itself and therefore do not need to be translated. Any kind of proprietary standard, such as *Alexa Presentation Language (APL)*, would increase complexity and decrease interoperability. In addition, this scenario involves two types of modalities: the voice interface and some type of visual interface such as touchscreens. Therefore, separate data formats should be used for the different interfaces. So for the generation of speech responses, *Speech Synthesis Markup Language (SSML)* seems to be suitable. To be more precise, this XML-based markup language will be supported by the Google Assistant [17], Alexa [18] and other assistants[1]. This format can be used to describe, how a VA should generate an auditory

[1]https://dueros.baidu.com/dbp (Accessed: 2022-04-11)

response, with regard to emphasis, pauses and pronunciation[2]. The visual responses, on the other hand, are based on typical web technologies such as HTML, JavaScript and CSS. Thus, if the VA's hardware has a graphical interface, the responses can be displayed as web applications. Moreover, the use of these technologies greatly improves the interoperability of the system. This is because the visual response can be interpreted using any type of web browser, regardless of which end device is used.

Once it has been determined how the data is to be formatted, the internal architecture structure must be designed. With regard to the generic claim of this approach, the actual business logic should be implemented as modular as possible in order to be easily extensible and changeable. We therefore propose to use a so-called microservice architecture pattern. Here, the server is structured as a combination of several encapsulated software components that interact via a clearly defined communication paradigm [19]. This allows certain components to be developed, maintained and extended independently, which in turn increases the maintainability and modularity of the system [19]. However, it should also be kept in mind that the use of this design pattern can also increase the complexity of the system. Especially if some components are distributed across multiple network nodes. Finally, the decision may also be influenced by the experience of the designer. Therefore, for the above reasons, we propose to use the microservice architecture pattern to maximize the modularity of the system.

Another important component of this architectural approach is the second server interface, through which all other requests to access patient data are handled. Therefore, it is mainly intended for professionals such as *Healthcare Providers (HCPs)* or scientists. It allows these users to access raw, unprocessed data. This data can then be further processed for specific questions or analyzed to gain new insights into certain diseases. However, it must always be remembered that this personalized health data should be treated with the utmost sensitivity. Finally, we also recommend implementing appropriate rights management that regulates all data access. This includes, for example, multiple user roles with different authorizations. To protect patient privacy, data should be anonymized prior to access. Ultimately, then, the main focus of this interface must be on providing easy and responsible access to well-protected data.

VI. Proof of Concept

After the elicitation of essential aspects of the underlying architecture, the following chapter deals with the prototypical implementation of our concept. In order to get a better understanding of the users' needs, the support of patients with therapeutic homework was chosen as a reference scenario. Furthermore, a selected *Voice Assistant (VA)*, Amazon Alexa, was utilized for the initial implementation. However, since this is a generic approach, Alexa-specific aspects and limitations are highlighted.

Since the server is considered the centerpiece of our architecture, this component was implemented first. It consists of our two previously designed interfaces, one handling voice interactions and the other allowing scientists to access patient data. Both interfaces, as well as the core of the architecture, are based on the JavaScript server framework NodeJS[3]. On the one hand, NodeJS is particularly suitable for dealing with the previously mentioned web technologies for visually supported response generation. On the other hand, both Google and Amazon already offer SDKs for their assistants, which are also available for NodeJS. Thus, both interfaces could easily be implemented using the same language, which in turn increases maintainability. Thanks to the microservice architecture pattern, further developments are flexibly possible. This represents a major advantage, especially when considering that data processing and evaluation could be done in a more suitable language such as Python. To this end, all components have to fulfill only one requirement to ensure communication: They all have to implement the same communication paradigm, which is based on the well-known REST principles.

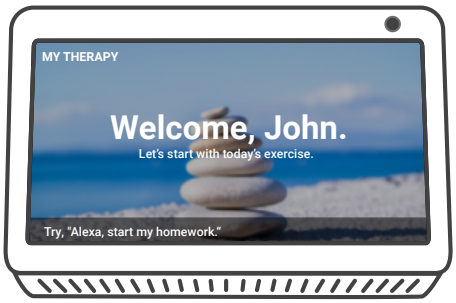

Fig. 3: Exemplary personalized visual VA frontend

After implementing the core functionalities on the server, the next important component of our presented architecture is the audio-visual VA front-end. As mentioned earlier, our proposed approach uses only established, standardized formats. Therefore, *Speech Synthesis Markup Language (SSML)* was used to implement the auditory part, while the visual components are based on HTML, CSS and JavaScript (see Figure 3). However, during the implementation it turned out that Google Assistant only supports a subset of the SSML specification [17]. Consequently, only a selection of SSML tags supported by both reference assistants could be used. Otherwise, the mediator would have to distinguish between the different vendors when generating an assistant's auditory response. This in turn would increase the complexity of our intermediate layer. The biggest challenge was therefore the definition of this SSML tag subset.

The last component to be presented here is the dashboard for professionals. This is a kind of reference implementation of a web front-end based on the server interface for the professionals mentioned previously. Since it is only a prototype, the security mechanisms described earlier have not

[2]https://www.w3.org/TR/speech-synthesis11/ (Accessed: 2022-04-11)

[3]https://nodejs.org/en/ (Accessed: 2022-04-11)

been implemented. What has been implemented, however, is a visual dashboard interface for therapists and other *Healthcare Providers (HCPs)*. This web application could be used, for example, to make ad-hoc adjustments to a patient's treatment plan, receive feedback that a user might give to their assistant, or monitor the progress of a therapy. The underlying server interface was also built on REST principles to meet the requirements of contemporary technologies. In addition, the web application was implemented using the JavaScript web framework ReactJS.

VII. Discussion

In the following section, we discuss how our proposed architectural design meets the requirements from Table I.

The POC has shown that the use of widely used data formats such as *Speech Synthesis Markup Language (SSML)* increases interoperability with different clients. However, due to the varying degrees of support for these formats, the off-the-shelf interoperability is not as far as we had anticipated. Therefore, this can be seen as another challenge for generalizing speech agent interactions.

Another important aspect of the architecture is the natural conversation flow described earlier. By using contextual information and historical data, we tried to make the user interactions as natural as possible. However, the use of interaction models made the conversations less flexible. Since this is a measure to reduce server complexity, this is considered a tradeoff. However, to enable a truly natural and personalized conversation, further development is needed.

The last important aspect of this discussion is the lack of transparency of the *Natural Language Understanding (NLU)* cloud services mentioned. This means that it is difficult to understand how a user's intention is detected. This raises questions about data protection and possible dependencies in particular.

VIII. Summary and Outlook

As seen in section VII, there are some non-negligible challenges that can arise in the generic integration of speech assistance into the therapeutic context. This paper therefore presented an approach for the corresponding technical implementation (see section V). It illustrates the complexity of designing, planning and implementing such a system. Furthermore, it is shown that generalizing speech interactions to understand the user's intentions and provide personalized healthcare support is one of the main challenges in such scenarios. Especially when considering the differences within the functional areas of the assistants. Finally, it was mentioned that the lack of transparency of current *Natural Language Processing (NLP)* cloud services increasingly raises questions regarding the protection of personal health data. However, as has been shown recently, vendors such as Amazon are already aware of this [20]. In summary, there is great potential for developments in the topic area covered.

References

[1] M. Schickler, R. Pryss, M. Stach, J. Schobel, W. Schlee, T. Probst, B. Langguth, and M. Reichert, "An it platform enabling remote therapeutic interventions," in *2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 2017, pp. 111–116. [Online]. Available: https://doi.org/10.1109/CBMS.2017.78

[2] Y. Liu, L. Wang, W. R. Kearns, L. Wagner, J. Raiti, Y. Wang, and W. Yuwen, *Integrating a Voice User Interface into a Virtual Therapy Platform*. New York, NY, USA: Association for Computing Machinery, 2021. [Online]. Available: https://doi.org/10.1145/3411763.3451595

[3] M. Schickler, R. Pryss, J. Schobel, and M. Reichert, "Supporting remote therapeutic interventions with mobile processes," in *2017 IEEE International Conference on AI Mobile Services (AIMS)*, 2017, pp. 30–37. [Online]. Available: https://doi.org/10.1109/AIMS.2017.13

[4] A. Ermolina and V. Tiberius, "Voice-controlled intelligent personal assistants in health care: International delphi study," *J Med Internet Res*, vol. 23, no. 4, p. e25312, Apr 2021. [Online]. Available: https://doi.org/10.2196/25312

[5] H. Chen, X. Liu, D. Yin, and J. Tang, "A survey on dialogue systems: Recent advances and new frontiers," *Acm Sigkdd Explorations Newsletter*, vol. 19, no. 2, pp. 25–35, 2017.

[6] (2022) Amazon alexa voice ai — alexa developer official site. [Online]. Available: https://developer.amazon.com/en-US/alexa

[7] (2022) Google assistant — google developers. [Online]. Available: https://developers.google.com/assistant

[8] R. Kraft, A. R. Idrees, L. Stenzel, T. Nguyen, M. Reichert, R. Pryss, and H. Baumeister, "esano–an ehealth platform for internet-and mobile-based interventions," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021, pp. 1997–2002.

[9] G. Vlaescu, A. Alasjö, A. Miloff, P. Carlbring, and G. Andersson, "Features and functionality of the iterapi platform for internet-based psychological treatment," *Internet Interventions*, vol. 6, pp. 107–114, 2016.

[10] D. Dojchinovski, A. Ilievski, and M. Gusev, "Interactive home healthcare system with integrated voice assistant," in *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, 2019, pp. 284–288.

[11] L. Qiu, B. Kanski, S. Doerksen, R. Winkels, K. H. Schmitz, and S. Abdullah, "Nurse amie: Using smart speakers to provide supportive care intervention for women with metastatic breast cancer," in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–7.

[12] K. H. Schmitz, X. Zhang, R. Winkels, E. Schleicher, K. Mathis, S. Doerksen, L. Cream, J. Rosenberg, R. Kass, M. Farnan, P. Halpin-Murphy, R. Suess, D. Zucker, and M. Hayes, "Developing "Nurse AMIE": A tablet-based supportive care intervention for women with metastatic breast cancer," *Psycho-Oncology*, vol. 29, no. 1, pp. 232–236, Jan. 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/pon.5301

[13] M. S. Broder, "Making optimal use of homework to enhance your therapeutic effectiveness," *Journal of rational-emotive and cognitive-behavior therapy*, vol. 18, no. 1, pp. 3–18, 2000. [Online]. Available: https://doi.org/10.1023/A:1007778719729

[14] F. Laricchia, "Global smart speaker market share 2021," Mar 2022. [Online]. Available: https://www.statista.com/statistics/792604/worldwide-smart-speaker-market-share/

[15] (2022) Create the interaction model for your skill. [Online]. Available: https://developer.amazon.com/en-US/docs/alexa/custom-skills/create-the-interaction-model-for-your-skill.html

[16] (2021, 08) Build conversation models. [Online]. Available: https://developers.google.com/assistant/conversational/build/conversation

[17] (2022, 03) Ssml (dialogflow). [Online]. Available: https://developers.google.com/assistant/df-asdk/ssml

[18] (2022) Speech synthesis markup language (ssml) reference. [Online]. Available: https://developer.amazon.com/en-US/docs/alexa/custom-skills/speech-synthesis-markup-language-ssml-reference.html

[19] I. Nadareishvili, R. Mitra, M. McLarty, and M. Amundsen, *Microservice architecture: aligning principles, practices, and culture*. O'Reilly Media, Inc., 2016.

[20] (2022) Voice for health and wellness. [Online]. Available: https://developer.amazon.com/en-US/alexa/alexa-skills-kit/get-deeper/custom-skills/healthcare-skills