SN Computer Science (2022) 3:267
https://doi.org/10.1007/s42979-022-01172-3

ORIGINAL RESEARCH

A Systematic Review of Voice Assistant Usability: An ISO 9241–11 Approach

Faruk Lawal Ibrahim Dutsinma[1] · Debajyoti Pal[1] · Suree Funilkul[2] · Jonathan H. Chan[1]

Received: 6 January 2022 / Accepted: 20 April 2022 / Published online: 3 May 2022

Abstract

Voice assistants (VA) are an emerging technology that have become an essential tool of the twenty-first century. The VA ease of access and use has resulted in high usability curiosity in voice assistants. Usability is an essential aspect of any emerging technology, with every technology having a standardized usability measure. Despite the high acceptance rate on the use of VA, to the best of our knowledge, not many studies were carried out on voice assistants' usability. We reviewed studies that used voice assistants for various tasks in this context. Our study highlighted the usability measures currently used for voice assistants. Moreover, our study also highlighted the independent variables used and their context of use. We employed the ISO 9241-11 framework as the measuring tool in our study. We highlighted voice assistant's usability measures currently used; both within the ISO 9241-11 framework, as well as outside of it to provide a comprehensive view. A range of diverse independent variables are identified that were used to measure usability. We also specified that the independent variables still not used to measure some usability experience. We currently concluded what was carried out on voice assistant usability measurement and what research gaps were present. We also examined if the ISO 9241-11 framework can be used as a standard measurement tool for voice assistants.

Keywords Voice assistants · Systematic literature review · Usability · User experience · ISO 9241-11 framework

Introduction

Voice assistants (VAs) which are also called intelligent personal assistants are computer programs capable of understanding and responding to users using synthetic voices. Voice assistants have been integrated into different technological devices, including smartphones and smart speakers [1]. The voice modality is the central mode of communication used by these devices, rendering the graphic user interface (GUI) inapplicable or less meaningful [2]. People use VA technology in different aspects of their lives, such as for simple tasks like getting the weather report [3] or managing emails [4]. In addition, the VA can perform complex tasks like client representative tasks [5] and controllers in autonomous vehicles [6]. In other words, VA's can revolutionize the way people interact with computing systems [7]. Currently, there is a massive global adoption of voice assistants. A report in [8] indicates that 4.2 billion VA's were adopted and used in 2020 alone, with a projected increase to 8.4 billion by 2024. The popularity of VA's has led to a greater research attention to its usability and user experience aspect.

Usability is a critical factor in the adoption of voice assistants [9]. A study by Zwakman et al. [10] highlighted the importance of usability in voice assistants [9]. An additional study by Coronado et al. [11] reiterated the importance of usability in human–computer interaction tools. Numerous studies have been carried out on the usability heuristics used in a VA, each study adopting a unique approach. A study by Maguire [12] used the Nielsen and Molich versions of Voice

✉ Debajyoti Pal
debajyoti.pal@mail.kmutt.ac.th; debajyoti.pal@gmail.com

Faruk Lawal Ibrahim Dutsinma
lawal.faruk@mail.kmutt.ac.th

Suree Funilkul
suree@sit.kmutt.ac.th

Jonathan H. Chan
jonathan@sit.kmutt.ac.th

[1] Innovative Cognitive Computing (IC2) Research Center, King Mongkut's University of Technology Thonburi, Bangkok, Thailand

[2] School of Information Technology, King Mongkut's University of Technology Thonburi, Bangkok, Thailand

User Interface (VUI), and the heuristic Voice User Interface (VUI), to evaluate the ease of use of the VA's. The study affirmed both the two heuristics were appropriate. However, the study noted that one was less problematic to use than the other [12]. A further study tested VUI heuristics to measure VA efficacy [13]. However, a critical factor that prevents the VA from adopting the heuristic currently available is the absence of a graphical user interface (GUI). Despite numerous studies on heuristics, the level of satisfaction is still low [14]. Furthermore, heuristics cannot be used as a standardized approach because they are approximate strategies or empirical rules for decision-making and problem-solving that do not ensure a correct solution. According to a study by Murad [16], the absence of standardized usability guidelines when developing VA interface presents a challenge in the development of an effective VA [15]. Another report from Budi & Leipheimer [17] also suggests that the usability of the VA's requires improvements and standardization [16]. To create a standard tool a globally recognized and well-known organization is critical in the process because it eliminates bias and promotes neutrality [17]. The International Organization for Standardization (ISO) 9241-11 framework is one of the standard usability frameworks widely used for measuring technology acceptance.

According to the ISO 9241–11 framework, usability is defined as "the degree to which a program may be utilized to achieve measurable objectives with effectiveness, efficiency, and satisfaction in a specific context of usage" [18]. ISO 9241-11 provides a framework for understanding and applying the concept of usability in an interactive system and environment [19]. The main advantage of using the ISO standard is that industries and developers do not need to build different design measurement tools. This standard is intended to create compatibility with new and existing technologies, and also create trust [20]. Currently, the system developers do not have any standardized tool created specifically for the measurement of VA usability, consequently, the measures are decentralized, causing confusion among developers. The lack of in-depth assessment of the current heuristics used in the VA design affects the trust and adaptability of their users [15]. Other emerging technologies such as virtual reality [21] and game design [22] have understood the importance of creating an acceptable standardized measurement tool when designing new interfaces. Therefore, VA technology could also benefit significantly from the same concept. As evident from the above discussion, there is little to no focus on VA standardization.

Our study presents a systematic literature review comprising works carried out on the usability of voice assistants. In addition, we use the ISO 9241-11 framework as a standardized measurement tool to analyze the findings from the studies we collected. We chose the ACM and IEEE databases for the selection of our articles because both contain a variety of studies dealing with the usability aspects of VA's. The following are the contributions of this literature review to the Human--Computer Interaction (HCI) community:

1. Our work highlights the studies currently carried out on VA usability. This includes the independent and dependent variables currently used.
2. Our study highlights the factors that affect the voice assistants' acceptance and impact the user's total experience.
3. We identify and explain some attributes unique to only voice assistants, such as machine voice.
4. We also highlight the evaluation techniques used in previous studies to measure usability.
5. Finally, our study tries to compare the existing usability studies with the ISO 9241-11 framework. The decentralized approach of the VA usability measurement makes it vague to understand if the ISO 9241-11 framework is being adhered to whilst developing the usability metrics.

We hope that our input will highlight the integration of the current existing VA usability measures with the ISO 9241-11 framework. This will also verify whether the ISO 9241-11 framework can serve as a standard measure of usability in voice assistants. In conclusion, our study tries to answer the following four research questions:

- RQ_1: Can the ISO 9241–11 framework be used to measure the usability of the VA's?
- RQ_2: What are the independent variables used when dealing with the usability of VA's?
- RQ_3: What current measures serve as the dependent variables when evaluating the usability of VA's?
- RQ_4: What is the relationship between the independent and dependent variables?

The remaining work is structured as follows. The second section presents the related work. This highlights what previous literature review studies had been carried out on voice agents' usability; furthermore, the section also highlights the emergent technology that employed the ISO 9241-11 framework as a usability measuring tool. This is followed by the methodology section, which presents the inclusion and exclusion criteria used together with the review protocol. Furthermore, the query created for the database search is presented, and the database to be used is also selected. The fourth section presents the result and analysis. In this phase, the article used for this study is listed. Also, the research questions are answered. The fifth section contains discussion on the result analysis. This includes a more detailed explanation of the relationships between independent and dependent variables. Our insights and observations are included in this section as well.

Literature Review

Previous Systematic Reviews

There have been a number of systematic literature reviews concerning VA's over the years. Table 1 presents the information for a few of the relevant works.

As highlighted in Table 1, multiple systematic literature reviews have been carried out on VA's usability over the years. However, each study has a specific limitation and gap for improvement. For instance, some studies focus on the usability of voice assistants used only in specified fields such as education [25] and health [36]. Other studies focus on the usability of voice assistants concerning only specific age groups, such as older adults [28]. Likewise, although an in-depth analysis of the usability of the VA's is carried out involving every usability measure in [32], this study does not use the ISO 9241 framework as a measuring standard. On the other hand, another study in [33] although uses the ISO 9241 framework as a measuring standard, however, the usage context was chatbots focusing primarily on text-based communication instead of voice. Overall, the available literature reviews on VA's usability listed in Table 1 supports the view that very few of the current literature review studies on VA's use the ISO 9241-11 framework as an in-depth tool for measuring usability.

The ISO 9242-11 Usability Framework

The ISO 9241-11 is a usability framework used to understand usability in situations where interactive systems are used and employed, which includes framework environments, products, and services [39]. Nigel et al. [40] conducted a study to revise the ISO 9241-11 framework standard, which reiterates the importance of the framework within the concept of usability. A number of studies have been conducted on various technologies using the ISO 9241–11 framework as a tool to measure their usability. This shows the diversified approach when using the framework. For instance, a study by Karima et al. (2016) proposed the use of ISO 9241-11 framework to measure the usability of mobile applications running on multiple operating systems by developers, in which the study identified display resolution and memory capacity as factors that affect the usability of using mobile applications [41]. Another study used the ISO 9241-11 framework to identify usability factors when developing e-government systems [42]. This study focused on the general aspect of e-Government system development and concluded the framework could be used as a usability guideline when developing a government portal. In addition, the ISO 9241-11 framework was also used to evaluate other available methods and tools. For instance, a study by Maria et al. [44] used the framework to evaluate existing tools used in the measurement of usability of software products and artifacts on the web. The study compared existing tools with the ISO 9241-11 measures for efficiency, effectiveness and satisfaction [43]. ISO 9241–11 framework has also been employed as a method of standardization tool in the geographic field [44], game therapy in dementia [45], and logistics [46]. Despite the ISO 9241-11 usability framework being utilized in different aspects of old and emergent technologies, it has not been used with a VA in the past.

Methods

We performed a systematic literature review is this study using the guidelines established by Barbara [47]. These guidelines have been widely used in other systematic review studies as a result of their rigor and inclusiveness [48]. In addition, we have added a new quality assessment process to our guidelines. The quality assessment is a list of questions that we use to independently measure each study to ensure its relevance for our review. Our quality evaluation checklists are derived from existing studies [49, 50]. The complete guidelines used in this section comprises of four different stages:

1. Inclusion and exclusion criteria
2. Search query
3. Database and article selection.
4. Quality assessment.

Inclusion and Exclusion Criteria

The inclusion and exclusion criteria used in our study are developed for completeness and avoidance of bias. The criteria we used for our study are:

a. Studies that focus on VA, with voice being the primary modality. In scenarios where the text or graphical user interfaces are involved, they should not be the primary focus.
b. Studies are only in the English language to avoid mistakes during translation from another language
c. The studies include at least one user and one voice assistant to ensure that the focus is on usability, not system performance.
d. Study has a comprehensive conclusion.

Table 1 Current literature reviews

#	Article name	Summary	Limitations	Usability focus
[23]	Smart Home Voice Assistants: A Literature Survey of User Privacy and Security Vulnerabilities	The study explores the potential use vulnerabilities encountered while using the voice assistant. The studies looked at the vulnerabilities, associated attack vectors, and possible mitigation measures that users can take to protect themselves during the use of voice assistant	Privacy and vulnerability are not the primary focus in usability	Personal Smart Home use
[24]	Intelligent personal assistants: A systematic literature review	The natural language interfaces allow the human–computer interaction by the translation of the human intention in the controls of the devices, the analysis of the speech or the gestures of the user. The article looked at the major trends, critical areas and challenges of an intelligent personal assistant. The study also proposed a taxonomy for IPA classification. The method used the population, intervention, comparison, outcome, and context (PICOC) criteria	The study did not conduct a thorough review of what was done with respect to the usability of the voice assistant	General use
[25]	Virtual Assistants for Learning: A Systematic Literature Review	The motivation, commitment and decreasing interest of students in the learning process has always existed, contributing to increased failures and dropouts. This can be attributed caused due to the difficulties with time management. The growing number of students in higher education makes it impossible to provide individual tutoring and support to each student. This paper systematically examines the use of virtual assistants in tertiary education It focuses on the technology which fuels them, their characteristics and their impact in the learning process	The Study focused on voice assistants used only within an educational environment that motivates users	Education
[26]	Voice-Based Conversational Agents for the Prevention and Management of Chronic and Mental Health Conditions: Systematic Literature Review	Chronic and mental diseases are increasingly prevalent throughout the world. As devices in our everyday lives offer more and more voice-based self-service, voice assistant can support the prevention and management of these conditions. This study highlights the current methods used in the evaluation of health interventions for the prevention and management of chronic and mental health conditions delivered through voice assistant	The study only focused on voice assistants used in the health environment alone	Health

Table 1 (continued)

#	Article name	Summary	Limitations	Usability focus
[27]	Tourists' Attitudes toward the Use of Artificially Intelligent (AI) Devices in Tourism Service Delivery: Moderating Role of Service Value Seeking	This study examines tourist attitudes towards the use of voice assistants in relatively more utilitarian or hedonic (air and hotel) tourism services. The results of the study suggest that tourism acceptance of VA is influenced by social influence, hedonistic motivation, anthropomorphism, expectation of performance and exertion, and emotions towards artificially intelligent devices. These results suggest that while the use of voice Assistants in the provision of functional services is acceptable, the use of AI devices in the delivery of hedonic services could backfire	The study was not on usability, to be specific, but on adopting voice assistants in the tourism environment	Tourism
[28]	Exploring How Older Adults Use a Smart Speaker–Based Voice Assistant in Their First Interactions: Qualitative Study Exploring How Older Adults Use a Smart Speaker–Based Voice Assistant in Their First Interactions: Qualitative Study	Smart speaker-based voice assistants promise support for the aging population, with the benefits of hands-free and eye-free interaction to process applications. This study explores how older adults experience and react to a voice assistant when they first interact with that person. The study discusses design implications that can positively influence older adults using voice assistant, including helping better understand how a voice assistant work and tailored to the needs of older adults	The Study group focused on was only older adults	General Use
[29]	A Meta-Analytical Review of Empirical Mobile Usability Studies	This document provides a usability assessment framework tailored to the context of a mobile IT environment. The study conducted a qualitative meta-analysis of more than 100 empirically based mobile usability studies. This study included the contextual factors studied, the dimensions of core and peripheral usage measured. Furthermore, open and unstructured tasks are under-utilized, and the effects of interaction between interactivity and complexity warrant further study	The impacts of User characteristics and environment on usability were not explored in the study	Mobile computing

Table 1 (continued)

#	Article name	Summary	Limitations	Usability focus
[30]	Evaluation of COVID-19 Information Provided by Digital Voice Assistants	Digital voice assistants are widely used to search for health information during COVID-19. With the rapidly changing nature of COVID-19 information, there is a need to assess the COVID-19 information provided by voice assistants to meet consumer needs and prevent disinformation. The goal of this study is to evaluate the COVID-19 information provided by voice assistants in terms of relevance, accuracy, usability and reliability. The study found that information about this pandemic is evolving rapidly and that users must use good judgment when obtaining COVID-19 information from voice assistants	The study focused only on voice assistants used in Covid-19 related issues	Health
[31]	The human side of human-Chatbot interaction: A systematic literature review of ten years of research on text-based chatbot	Over the last ten years there has been a growing interest around text-based chatbot, software applications interacting with humans using natural written language. However, despite the enthusiastic market predictions, 'conversing' with this kind of agents seems to raise issues that go beyond their current technological limitations, directly involving the human side of the interaction. This study suggests a number of research opportunities that could be explored over the next few years	The study focused only on chatbot with textual modality. Moreover, chatbot use a Graphic User Interface that is not present in voice assistants	General Use
[32]	Voice in Human–Agent Interaction: A Survey	Social robots, conversational agents, voice assistants and other embodied AIs are increasingly a characteristic of daily life. The connection between these different types of intelligent agents is their ability to interact with people by voice. The voice becomes an essential mode of embodiment, communication and interaction between IT operators and end users. This study presents a meta-synthesis of the voice of agents in the conception and experience of agents from a man-centered point of view: voice assistant	The study did not use the ISO 9241–11 framework as a reference in their measurement scale	General Use
[33]	Usability of Chabot's: A Systematic Mapping Study	The use of chatbot has increased considerably in recent years. As a result, it is essential to integrate conviviality into their development. For this reason, it is essential to integrate conviviality in their development. The study identifies the state of the art in the conviviality of chatbot and the applied techniques of human–computer interaction, to analyze how to assess the conviviality of chatbot	The study focused only on chatbot with textual modality. Moreover, chatbot use a Graphic User Interface that is not present in voice assistants	General Use

Table 1 (continued)

#	Article name	Summary	Limitations	Usability focus
[34]	A Literature Review On Chatbot In Healthcare Domain	The study highlighted Chabot used in the healthcare environment. Also, it compares the techniques such as NLU, NLG, and ML used in chatbot development	The study deals with chatbot with textual modality. Also, the study deals with chatbot used only in a healthy environment	health
[35]	Review of Chatbot Design Techniques	The study reviewed the techniques and factors considered when designing a chatbot.Also it highlighted how chatbot worked and what are the type of approaches that are available for chatbot development	The study focuses on chatbot with textual modality, which is different from a voice assistant	Commerce
[36]	A Systematic Literature Review of Medical Chatbot Research from a Behavior Change Perspective	The study examined the literature on how people feel about using a medical chatbot in medical communication services. Moreover, The study recommended five design-orientation and highlighted the behavioral aspects such as acceptance, usage, and effectiveness when using chatbot	The study focuses on chatbot with textual modality, which has different factors than voice assistants	Health
[37]	A review of chatbot in education: Practical steps forward	The study focused on Chatbot applied within an educational environment; it highlighted how Chatbot are currently being used in a broader educational environment. Moreover, the study also recommended how Chatbot can be applied to enhance students learning experience	The study focuses on chatbot, with textual modality, with different factors than voice assistants	Education
[38]	Human-like communication in conversational agents: a literature review and research agenda	The study identified the voice assistant human-like behaviors that have the most effect on relational outcomes during communication	An in-depth analysis of user and conversation assistants attributes was not carried out. Moreover, the study only focused on voice assistants used in management alone	Management

e. Released between 2000 and 2021, because during this period the vocal assistants started to gain notable popularity

The exclusion criteria are:

a. Studies with poor research design, where the study's purpose is not clear are excluded.
b. White papers, posters, and academic thesis are excluded.

Search Query

We created the search query for our study using keywords arranged to search the relevant databases. We went through previous studies to find the most relevant search keyword to find what is commonly used in usability studies. After numerous debates among the researchers and seeking two HCI expert's opinion, we chose the following set of keywords: usability, user experience, voice assistants, personal assistants, conversational agents, Google Assistant, Alexa, and Siri. We connected the keywords with logical operators (AND and OR) to yield accurate results. The final search string used was ("usability" OR "user experience ") AND ("voice assistants" OR "personal assistants" OR "conversational Agents" OR "Google Assistant" OR" Alexa" OR "Siri"). The search was limited to the abstract and title of the study.

Database and Article Selection

Figure 1 highlights the graphic presentation of the selection and filtering process. The figure is adapted from the Prisma flow diagram [51]. As earlier stated, two databases are used as the sources for our article selection: the Association for Computing Machinery (ACM) and the Institute of Electrical and Electronics Engineers (IEEE). Both databases we used in our study contain the most advanced studies on VA and are highly recognized among the HCI community. The search query returned 340 results from the ACM database and 280 results from the IEEE database. 720 items in both databases were checked for duplication and 165 documents (23%) were found to be duplicated and hence removed. Additionally, more items were filtered by title and abstract. We utilized keyword match to search the title; however, the abstract was read to identify the eligibility criteria. In addition, 399 documents (72%) were removed because they did not meet the eligibility criteria. Finally, 121 documents were removed that were not consistent with the research objectives of our study. At the end of the screening process 29 articles (19%) were finally included in this literature review.

Quality Assessment

The selected items presented in Table 2 are used for assessing the quality of the selected articles. The process was deployed to ensure the reported contents fit into our research. The sections collected from articles such as the methodology used, analysis done, and the context of use within each article were vital to our study. Each question is a three-point scale: "Yes" is scored as 1 point, which means the question is fully answerable. "Partial" is scored as 0.5, which means the question is vaguely answered, and "NO "is scored as 0, which means it is not answered at all. All the 29 sets of finally included articles passed the quality assessment phase.

Result and Analysis

List of Articles

This section lists and discusses the articles collected in the previous stage. Table 3 presents the list of all the

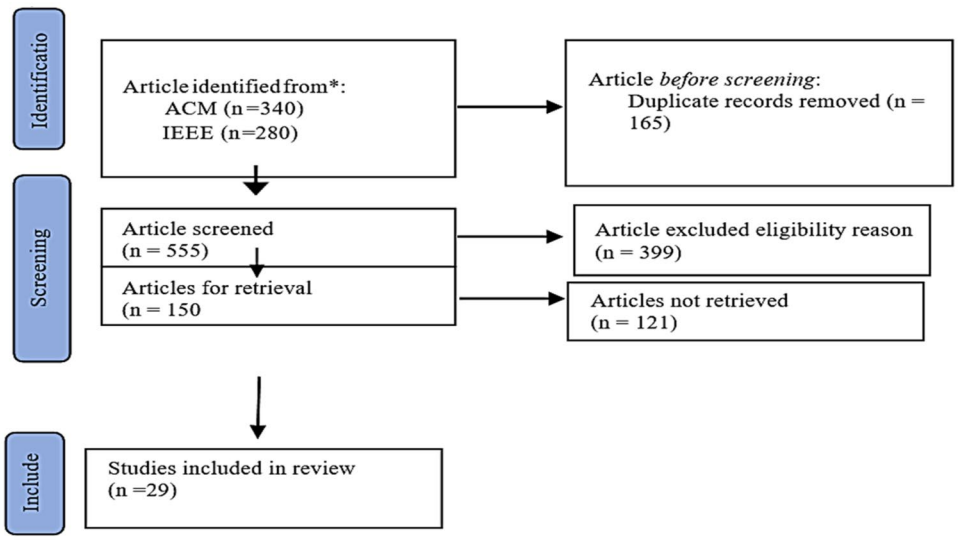

Fig. 1 Article selection process

Table 2 Quality assessment checklist

Checklist	Definition
C_1	Are the study aims and objectives clearly stated
C_2	Is the article well designed to achieve these aims?
C_3	Are the independent variable in the study clearly defined?
C_4	Are the dependent variable in the study clearly defined?
C_5	Is the study discipline stated clearly?
C_6	Are the data collection methods clearly stated
C_7	Does the study explain the reliability and validity of the measures?
C_8	Are the analysis techniques described adequately?
C_9	Are the users/participants' numbers stated clearly?
C_{10}	Do the results add to the literature?

compiled articles. Moreover, we identified the usability focus of each study.

Voice Assistant Usability Timeline

We grouped the collected research into three categories, each representing a range of time frames (Fig. 2). The categorization is based on voice assistant period breakthroughs. The first category is from 2000 to 2006, which was the year of social media and camera phones, also known as the year of the Y2K bug in telecommunications. During these years, conversational agents started to get noticed with the introduction of the inventions such as the Honda's Advanced Step in Innovative Mobility (ASIMO) humanoid robot [80]. The second category ranges from 2007 to 2014. During these years technological advancements got users more exposed to voice assistants through embedding them into smartphones and computers. For instance, Apple first introduced SIRI in 2011 [81], and Microsoft introduced Cortana in 2014. The last category ranges from 2015 to 2021. This was when the massive adoption of voice assistants took place, making it an all-time high.

Based on the year of publication of our selected articles, Fig. 2 clearly shows that the study on VA's has expanded significantly over the last six years (2014–2021). This can be attributed to the invention of a smart speaker and phone with built-in voice agents [82]. Another reason for VA popularity is the COVID -19 outbreak that has given a fresh impetus towards touchless interaction technologies like voice [83].

Different Embodiment Types of VA's

Smart speakers are the mostly used embodiment of VA's used in our selected articles. This is due to the current popularity of commercial smart speakers such as Alexa, HomePod, etc. A 2019 study showed that 35% of US households have an intelligent smart speaker, and projected to reach 75% by 2025 [84]. Use of humanoids is also popular because usability measures such as anthropomorphism are essential for voice assistant usability [85]. Furthermore, Fig. 3 shows that only a few studies were done on car interface voice assistants. Car interfaces are vocal assistants that act as intermediaries between the driver and the car. The VA car interface allows drivers to access car information and also be able to perform the task without losing focus on driving. The fourth type of software interface refers to a voice assistant software embedded inside smartphones or computers. The studies we have collected have used either the commercialized form of the software interface, such as Alexa and Siri, while others have developed new voice interfaces that are easily accessible to users due to the adoption of smartphones and computers assistants using programming codes and skills. Nevertheless, both are in the forms of different software agents.

Component of ISO 9241-11 Framework

The ISO 9241-11 framework highlights two components, the context of use and usability measure [18]. We concentrate on both components to highlight any correlations between usability metrics and the context of use in the selected articles. The context of use consists of the different independent variables along with the techniques used for analyzing them. Likewise, the usability measure represents the dependent variables, i.e., the effect that the independent variables have on the overall experience of the users. Accordingly, the analysis is presented in a bi-dimensional manner in the following sections.

Context of Use

Independent Variable We split the context of use into an independent variable and the techniques used. The inde-

Table 3 List of compiled articles

#	Article name	Voice assistant type	Usability measure	Years
1	An Exploration of Speech-Based Productivity Support in the Car [52]	Car Interface	Effectiveness	2019
2	Exploring Effects of Conversational Fillers on User Perception of Conversational Agents [53]	Smart Speaker	Effectiveness, machine Voice(perceived intelligence)	2019
3	I Almost Fell in Love with a Machine": Speaking with Computer Affects Self-disclosure [54]	Software Interface	Trust	2019
4	Clarifying False Memories in Voice-based Search [55]	Smart Speaker	Satisfaction, efficiency, cognitive load	2019
5	The Effects of Anthropomorphism and Non-verbal Social Behavior in Virtual Assistants [56]	Smart Speaker Humanoid	machine Voice(perceived humanness, social presence), cognitive load(attention)	2019
6	An End-to-End Conversational Style Matching Agent [57]	Smart Speaker	Trust	2019
7	Tandem Track: Shaping Consistent Exercise Experience by Complementing a Mobile App with a Smart Speaker [58]	Smart Speaker Software Interface	Efficiency, Effectiveness	2020
8	Mapping Perceptions of Humanness in Intelligent Personal Assistant Interaction [59]	Smart speaker Software Interface	Machine voice(perceived Humanness), Effectiveness	2019
9	Pattern of Gaze in Speech Agent Interaction [60]	Humanoid	Machine voice (Social presence), cognitive workload	2019
10	Conversational Interfaces for a Smart Campus: A Case Study [61]	Smart Speaker Software Interface	Effectivity	2020
11	Mental Workload and Language Production in Non-Native Speaker IPA Interaction [62]	Smart Speaker Software Interface	Cognitive Load, Satisfaction	2020
12	User Experience of Alexa when controlling music – comparison of face and construct validity of four questionnaires [63]	Smart speaker	User satisfaction	2020
13	Machine Body Language: Expressing a Smart Speaker's Activity with Intelligible Physical Motion [64]	Humanoid	Machine Voice (Perceived humanness)	2020
14	Measuring the anthropomorphism, animacy, likeability perceived intelligence and perceived safety of robots [65]	Humanoid	Machine Voice (Perceived humanness, Anthropomorphism)	2008
15	At Your Service: Designing Voice Assistant Personalities to Improve Automotive [66]	Car interface	Attitude(Likeability, acceptance)	2019
16	Hey, Siri", "Ok, Google", "Alexa". Acceptance- Relevant Factors of Virtual Voice-Assistants [67]	Smart Speaker Software Interface	Attitude (Trust acceptance)	2019
17	User experience with smart voice assistants: the accent perspective [68]	Smart Speaker Software Interface	User satisfaction	2019
18	Empathy is all you need: How a conversational agent should respond to verbal abuse [69]	Software Interface	Effective, User satisfaction,Machine voice (social presence)	2020
19	Gendered Voice and Robot Entities: Perceptions and Reactions of Male and Female Subjects [70]	Humanoid	User satisfaction, attitude, effectiveness	2009
20	What If Conversational Agents Became Invisible? Comparing Users' Mental Models According to Physical Entity of AI Speaker [71]	Smart speaker	Attitude(trust), machine Voice(Anthropomorphism)	2020
21	Similarity is more important than expertise: Accent effects in speech interfaces [72]	Smart Speaker	Effectiveness, attitude(Trust), Efficiency, satisfaction	2007
22	Can Computer-Generated Speech Have Gender? An Experimental Test of Gender Stereotype [73]	Software Interface	Attitude, User satisfaction	2020

Table 3 (continued)

#	Article name	Voice assistant type	Usability measure	Years
23	Designing Social Presence of Social Actors in Human Computer Interaction [74]	Software Interface	satisfaction	2003
24	Improving Automotive Safety by Pairing Driver Emotion and Car Voice Emotion [74]	Software Interface	Effectiveness, Efficiency	2005
25	Designing Emotional Expressions of Conversational States for Voice Assistants: Modality and Engagement [75]	Humanoid	Cognitive load, Attitude	2018
26	The Use of Voice Input to Induce Human Communication with Banking Chabot's [76]	Smart Speaker	Attitude	2018
27	Face Value? Exploring the Effects of Embodiment for a Group Facilitation Agent [77]	Smart Speaker Software Interface	Attitude	2018
28	Trust in artificial voices: A "congruency effect" of first impressions and behavioral experience [78]	Humanoid	Attitude	2018
29	Children Asking Questions: Speech Interface Reformulations [79]	Smart Speaker	Cognitive load	2018

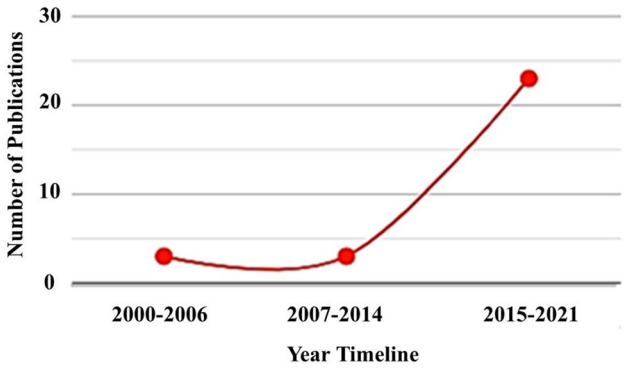

Fig. 2 Year of publication of selected articles

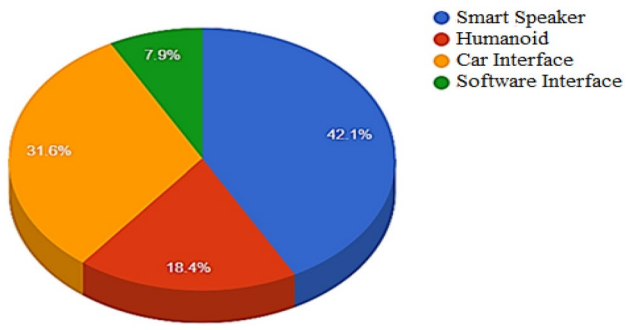

Fig. 3 Embodiment of Voice assistant used in selected studies

pendent variables presented in our study are the physical and mental attributes used to measure a given user interaction outcome. Furthermore, our study grouped the independent variables into five main categories. The grouping is shown in Fig. 4 and is based on the similar themes identified from the collected studies. The five groups included people (user attributes), voice (voice assistant attributes), task, conversational style, and anthropomorphic cues. The voice and people categories are the oldest independent variables used to measure usability. Their relevance is also seen in the recent studies, which indicate that researchers have a high interest in correlating users with the VA's. On the other hand, anthropomorphic clues and conversational styles are relatively new to the measurement of usability. The task-independent variable is the most used variable of late, perhaps because users always test the VA's ability to perform certain tasks. It also indicates that VA's are widely used for various functional and utilitarian aspects. The anthropomorphic cues are seldom used in the second phase (2007–2014). However, it is most widely used in the last range (2015–2021).

In Table 4 we highlight more details with regards to the different groups of the independent variable collected, and also present examples of the independent variables for each category. We highlight how the independent variables have been applied by the previous studies and in which environment they have been used. We defined each independent variable category in Table 4, and explained their sub-categories as well. As evident from Table 4, different independent variables are used together in multiple studies. For example, independent voice variables and independent people variables are used simultaneously in various studies, such as personality, gender, and accent. Similarities between multiple independent variables aid to understand the relationship between the variables themselves and their relationship with the usability measures. Furthermore, the table also highlights the kind of experiments carried out. Controlled experiments are effective methods for understanding the immediate cause and effect between variables. However, a noticeable drawback

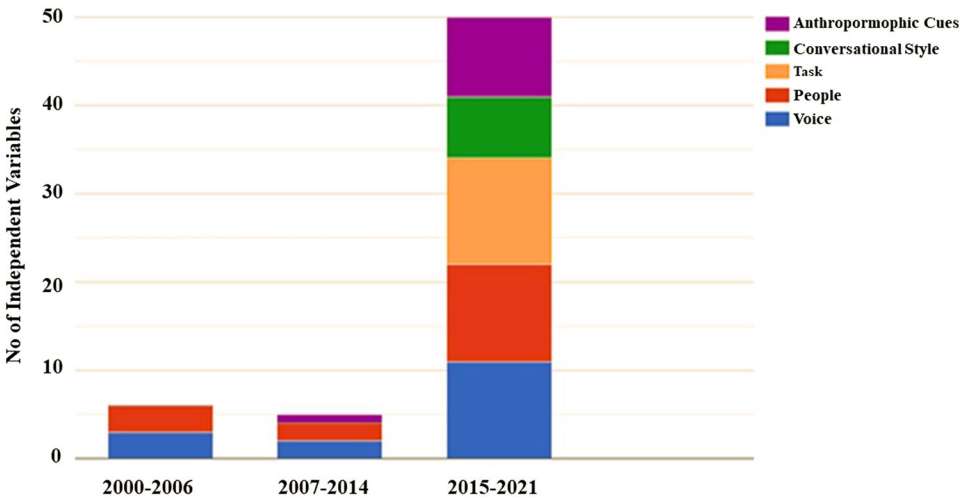

Fig. 4 Categories of independent variable use over the years

of controlled experiments is the absence of external validity. The results might not be the same when applied in real-world settings. For instance, the simulation experiment on cars is a controlled environment, a driver has no control over the domain in real life. The usability experience of the driver might be different in natural settings and that might sometimes prove fatal.

Techniques Used We identified seven techniques that researchers have used as shown in Fig. 5. The quantitative experiments are the most used and the oldest technique used on voice assistants based on our data collected. The quantitative method is sometimes used as a standalone experiment and sometimes with other techniques [54]. It is worthy of notice that cars simulation experiments involving VA's were first used in 2000. Other experiments on human communication with self-driving cars have been carried out since 1990's. making it one of the oldest techniques for usability measurement. More accurate technique was introduced later, such as the interaction design. The interaction design employed by studies such as [61] provides a real-time experiment scenario. This avoids the drawback such as bias when using quantitative methods. Factorial design studies are majorly used by studies that compare two or more entities in a case study [55]. They are utilized mainly by studies using two or more independent variables together.

Usability Measure (Dependent Variable)

This subsection of our study focuses on the usability measurement of our research. Moreover, the findings are used to answer RQ_1 and RQ_3. The ISO 9241-11 framework grouped usability measures into three categories; effectiveness, efficiency, and satisfaction. According to the ISO 9241-11 framework, "effectiveness is the accuracy and completeness with which users achieve specified goals.", Whereas "Efficiency is the resources expended concerning accuracy and completeness in which users achieve goals" and "satisfaction is the freedom from discomfort and positive attitudes towards the use of the product" [18].

In numerous studies, the usability measures used were clearly outside the scope of the ISO 9241-11 framework. In total, we identified three additional usability categories attitude, machine voice (anthropomorphism), and cognitive load. The graphical representation of the different usability measures identified in this study is presented in Figs. 6 and 7. Futhermore, the figures also highlights the percentage of studies that used the mentioned usability measures in the ISO 9241-11 framework and those that are outside the framework. Based on our compiled result, the user satisfaction and effectiveness are the earliest usability measures used when measuring VA's usability. Some studies used performance and productivity as subthemes to measure effectiveness [62]. The measure of usability has been carried out both subjectively and objectively. For instance, studies have measured the VA effectiveness by subjective means by using quantitative methods such as questionnaire tools [72]. In contrast, other studies have used objective methods such as average completed interaction [69]. Multiple usability measures are sometimes applied in the same research; for instance some studies measured effectiveness alongside efficiency and satisfaction [66, 70]. Learnability, optimization, and ease of use have been used as subthemes to measure efficiency. Interactive design is the most effective experiment that provides real-time results employed [56, 79]. The ISO 9241-11 framework works well with effectiveness, efficiency, and satisfaction; however, the users have more expectations from the voice assistant with the recent advancement of VA capabilities. Our compiled result showed that more than half of the studies are not carried out in accordance with the standard ISO 9241-11 framework

Table 4 Independent variables and their categorization

Category definition	Independent variable	Instances	Applications	Environment
Voice The voice category comprised of independent variables that are associated with the voice assistants, these are attributes that the voice assistants possess)	Voice personalities	(Energetic vs Subdued), (Introvert and extrovert)	A study Paired the driver's emotions with that of the Car Voice Emotion state (Energetic and Subdued) to test the effectiveness of similarity between voice and user personality [68]. Another study showed that a voice personality that uses a similar personality like the user creates more social presence [74]	Simulation Experiment, Controlled Environment,
	Voice gender	Male vs Female	Studies compared different gender voices (Male and Female) to measure social interaction and trust. Studies showed male voice has a more dominating effect on users than female voice [54, 70, 74]	Controlled Environment, Free real environment
	Voice Accent	Standard Southern British English accent VS Liverpool accent Vs Birmingham accent Vs synthetic voice, American Accent vs Swedish Accent, Native English speaker vs non-English speaker	Participants create trust expectancy based on the voice accent. The participants tend to trust information with a similar accent, more knowledgeable, sophisticated voice Assistants [68][68]	Controlled Environment,
People The people category comprised of independent variables that are associated with the users, these are attributes that the users can have	People gender	Male vs Female	Studies showed that Males and females view voice assistants differently in different aspects. Both genders have different takes in the form of embodiment of the voice assistants. Moreover, women trust voice assistants with a female voice. However, in a situation where there is a need for convincing the male voice assistants is more efficacious [67]	Controlled Environment, Free real environment, mixed environment
	Personality	Introvert vs Extrovert, happy vs upset	A study showed that a person's emotional state or personality could be affected the personality of the voice assistant. [74] A study showed that social presence is created when a person uses a voice assistant with a similar personality [54]	Simulation Experiment, Controlled Environment,

Table 4 (continued)

Category definition	Independent variable	Instances	Applications	Environment
	Query expression	Abuse (Insult, Threat, Swearing)	A study instructed the user to insult the voice assistants while communicating with it, and the VA's response affected the user's outlook and involved usability [69]	Controlled Environment
	Experience	UX metric, Self-Efficacy	The study Measured the user face validity and construct validity by correlating UX scores of questionnaires with each other. Another study shows that Participant self-efficacy and experience affect the trust, privacy and language performance of the Voice Assistant [74]	Controlled Environment, survey
	Voice accent	American Accent vs Swedish Accent, Native English speaker vs non-English speaker	Participants tend to trust information with a similar accent, then more knowledgeable content, English native speakers do exhaust more mental models when interacting with voice assistants [52, 68, 72]	Controlled Environment, Free real environment, mixed environment
Task characteristic The Task characteristic Comprised of independent variables that are associated with tasks that it's expected the user to carry out during the interaction, this also include the modality of the task	Modality	Voice mode, Textual mode, VA Facial Expression mode. (Smiley) Mixed Interface	A study used modality to test the social presence of the VA. The study shows participants feel a strong social presence when textual modality personality matches the voice personality. Another study showed that nonverbal emotional expressions such as Text box movement and VA Facial Expression mode (Smiley) affect user engagement [57]	emotional expression design experiment interactive task, controlled Environment,
	Context	Interactive Task, Drawing Task, Executable Task, Driving simulation task, auditory Task Controlling device Volume, audio speech to text	A study used the speech to text as a task on users during driving. The study measured driver engagement and concentration during driving. [52]Another study used the game theory concept on the users and asked the users to trust the VA in an investment scheme, where the users have a different opinion on what VA to trust [76]	emotional expression design experiment interactive task, controlled Environment, free real life environment, simulation

Table 4 (continued)

Category definition	Independent variable	Instances	Applications	Environment
Conversational Style This is the nature of the conversation from either the user during query or the response of the voice assistants	Response type	Empathetic (Avoidance vs Empathy vs Counterattack) Clarifying Query (No modification vs direct Modification vs negatively clarified) Conversational Fillers ("um",huh, uh)	The VA response affects the user usability experience, and A Study showed that When VA are insulted, their response type affects the participant's emotional engagement and attitude [69]. Another study showed that when VA has more information on a query, the follow-up question affects user engagement and efficiency[55] A study showed that Conversational fillers increase social interaction with the voice Assistant [53]	Controlled Environment,
	Communication Form	High Consideration(indirect) VS High Involvement (direct)	A study used participant High Consideration and High involvement linguistic style to realize. It is effective when used with a similar voice assistant's linguistic style. [57]	Controlled Environment, Mixed Environment
Anthropomorphic cues These are independent variables on voice assistants that exhibit human attributes and intelligence, this make the user perceive the voice assistants as human	Speech agents' Personification vs Speech agent personalization		A study compared VA personation, personalization, and neither to measure users' trust and engagement when used by children and adults. The result showed the personalized VA has the highest concentration and trust [79]	Controlled Environment,
	Embodiment type (audio Vs smart speaker), or (gaze vs no gaze), humanoid robot, Smart Speaker vs Anthropomorphic Robot (AMR). vs The Anthropomorphic Social Robot (AMSR)		Numerous studies have used embodiment type to measure usability; a study compared a VA with gaze with another VA without gaze to measure the user anthropomorphism. Another study compared physical smart speakers with the absence of speakers but just voice to test the user trust and engagement [56, 60, 67, 78]	Controlled Environment,

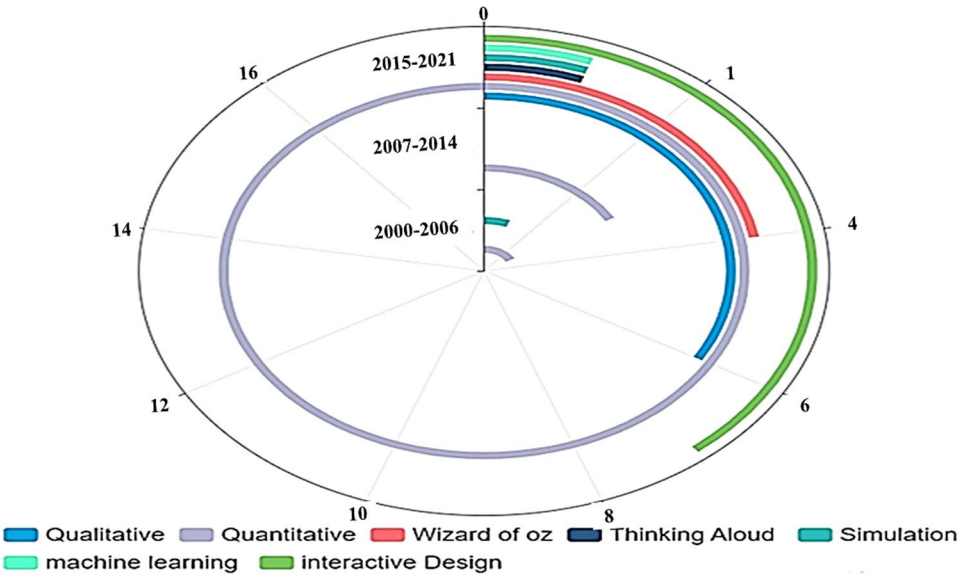

Fig. 5 Technique used in our studies over the span period of time

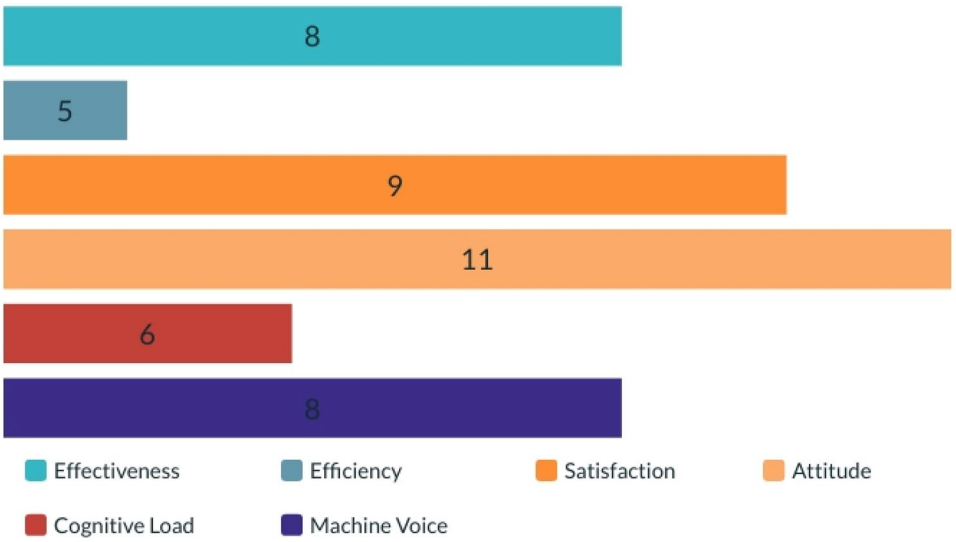

Fig. 6 Usability measurement used over the years on our compiled articles

(Fig. 7). The other usability measures we identified outside the ISO 9241-11 framework are attitude, machine voice, and cognitive load.

Attitude is a set of emotions, beliefs, and behavior towards the voice assistants. Attitude results from a person's experience and can influence user behavior. Attitude is subjected to change and is not constant. Understanding the user attitude towards the VA has become an active research area. Numerous studies have used different methods to measure subthemes of attitude such as trust, closeness, disclosure, smartness, and honesty [60, 78]. Likeability is also a subtheme of attitude, and it has been used to measure the compatibility, trust, and strength between the user and VA's [56,

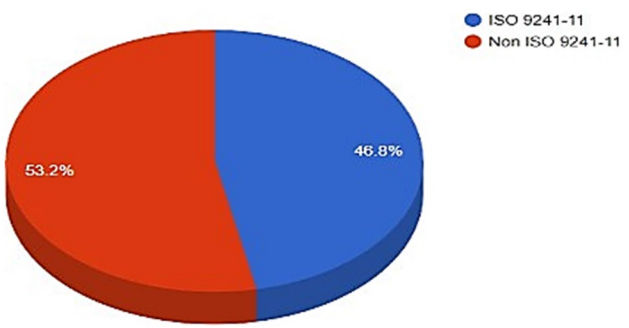

Fig. 7 Percentage of ISO 9241–11 framework usability measures and non ISO 9241–11

57]. Moreover, embodiment type affects the user attitude as well, A study highlighted how gaze affects the user attitude toward VA [59], and it shows VA with gaze creates trust.

We defined machine voice (anthropomorphism) as the user attribution of human characteristics and human similarity to the voice assistant. We considered machine voice an important usability measure that only applies to voice assistants due to their primary modality being the voice. Considering that fact, the measure of machine voice has also spiked currently it becomes obvious that it has been drawing a lot of interest. One of the direct purposes of the VA is to sound as humanly as possible. When the users will perceived the machines to be more human, it built more trust, which will result in a better usability experience.

The cognitive load might be mistaken for efficiency. Nevertheless, they are different. We defined cognitive load as the amount of mental capacity a person applies to communicate successfully with the VA. When it comes to VA, actions such as giving out commands require cognitive thinking and approach. The cognitive load is measured by specific characteristics unique to the VA, such as attention time during the use of the VA [76] and the user's mental workload during use [77].

To answer RQ_1 (*can the ISO 9241–11 framework be used to measure the usability of the VA's?),* none of the existing works have used the ISO 9241-11 framework solely for the purpose of usability evaluation. It has been supplemented by other factors that we have presented above that are outside the scope of this framework.

Relationship Between the Independent variables and Usability Measures

After identifying the independent and dependent variables, in Tables 5 and 6 we show how they are inter-related for having a better understanding of the usability scenario of the VA's. While Table 5 focuses on the ISO 9241-11 specific factors, Table 6 considers the non-ISO factors specifically.

The independent variables are grouped into categories and represented by table rows, with every category consisting of multiple independent variables. Moreover, the usability measures have been presented in the column of the table. Every usability measure is made up of different sub-themes, which are all presented on the table as well. The tables highlight the relationship between the independent and usability measures. An "X" mark present in each cell represents a study present between that independent variable and usability measure subtheme. Nonetheless, an empty cell indicates that there is no study carried out to link that relationship between the usability measure and independent variable.

Discussion

Independent Variable and Usability Measures

Our study revealed what has been previously carried out in VA usability and revealed the gaps that are yet to be addressed. We analyzed the usability measures and their relationship to the so-called independent variables. There is an easy accessibility to VA's due to the development of different embodiment types such as speakers, humanoids, and robots. However, there is so much less focus on embodiment types and their relationship to effectiveness and anthropomorphism, which needs more attention. Some relationship gaps and associations are apparent, while some are vague. For instance, the independent variable "*accent*", has often been connected with its effectiveness on users. However, what is left unanswered is if the VA accents impart the same efficacy on users of the same or different genders. Another notable gap is gender and efficiency, with very few studies on that. This will be an essential aspect to understand and apply with the recent massive adoption of voice assistants in different contexts. Another obvious gap is the query expression relationship with any ISO 9241-11 framework measures. The query expression is how a user expresses their query to the voice assistants. The query expression has been known to increase the trust and attitude of the user towards the VA. However, its relationship to usability measures such as efficiency, satisfaction, and effectiveness is still under-researched. Knowing the right way to ask queries (questions) defines the type of response a user gets. An incorrect response will be received if the right question is expressed incorrectly. From a mental model, when a user has too much energy and thought to frame a question, it affects the VA efficiency and satisfaction. However, this has not been proven by any current study.

The VA response types increase effectiveness and trust. However, its relationship to user acceptance is still unknown. Another exciting intersection is the anthropomorphic cues and attitude, which results from anthropomorphic emotional response than a practical one. Attitude is an emotional response to a giving state, hence its strong connection with anthropomorphism. The attitude toward the VA is a highly researched area [86]. Trust, likeability, and acceptance are subthemes that focused on the attitude usability measure. This can be attributed to the importance of trust while using emergent technologies such as voice assistants. User trust in voice assistants is an essential aspect with the rise of IoT devices, and user mistrust affects the acceptance and effectiveness of the VA's [87]. Multiple studies measured user trust while using machine voice categories as an independent variable. That could be attributed to the lack of GUI in VA. Furthermore, the voice modality must be enough

Table 5 Relationship between independent variables and ISO 9241–11 framework measurement

Dependent variables	Effectivity			Efficiency					Satisfaction		
Independent variables	Productivity	Performance	Value	Learnability	Optimization	Ease of use	Feasibility	Decision making	User experience	Continued use	Conformity
Voice Assistants											
Personalities	x	x	x			x			x		
Gender		x	x			x			x		x
Accent		x	x	x		x			x		
People											
Gender		x	x						x		x
Personality		x				x			x		
Query expression											
Experience				x		x					x
Accent	x	x	x	x		x			x		
Task characteristics											
Modility	x	x		x	x		x		x		x
Context							x		x		
Communication type											
Response type		x			x	x			x		x
Conversational type		x		x		x	x		x		x
Anthropomorphism											
Embodiement type		x		x			x	x	x	x	
Humanoid/robott		x							x		
Smart Speaker, Robot,Anthropormorphic Robot			x	x	x	x		x	x		

Table 6 Relationship between independent variables and non- ISO 9241–11 framework measurement

Dependent variables	Attitude			Machine voice			Cognitive load	
Independent variables	Trust	Likeability	Acceptance	Perceived intelligence	Perceived humanness	Social presence	Mental workload	Attention
Voice assistant								
Personalities	x	x			x	x		
Gender	x	x		x	x	x		
Accent	x	x		x	x	x	x	
People								
Gender	x			x		x		
Personality	x	x		x	x	x		
Query expression	x	x						
Experience	x	x					x	x
Accent	x	x		x	x	x	x	
Task Characteristics								
Modality	x	x	x	x	x	x	x	
Context	x			x	x	x		
Communication Type								
Response type	x	x		x	x	x	x	
Conversational type	x	x		x	x	x	x	
Anthropomorphism								
Embodiment type	x	x		x	x	x	x	x
Humanoid/robot	x	x			x	x		x
Smart speaker, robot, anthropomorphic robot	x	x					x	x

to cultivate user trust. Noticeably subjective methods were widely employed when measuring the user attitudes; even though subjective measures often relate to the variables they are intended to capture; however, they are also affected by cognitive biases.

The ISO 9241-11 framework is an effective tool when measuring effectiveness, efficiency, and satisfaction. However, it is not applicable when measuring usability's, such as attitude, machine voice, and mental load. These are all measurements that are uniquely associated with voice assistants. Therefore, the ISO 9241-11 framework could be expanded to include such usability aspects.

Technique Employed

The factorial design adapts well when used in a matched subject design experiments [56]. Based on the studies collected, machine learning is not well used as an analytic tool in usability. This could be attributed to the technical aspects of machine learning and it is still relatively a new field. However, with machine learning third-party tools more analysis will be carried out. Wizard of Oz, and interactive design started gaining popularity in 2015–2021. Moreover, the Wizard of Oz and interactive techniques are more effective when using independent variables such as anthropomorphic cues. The anthropomorphic cue independent variables is used with Wizard of Oz. techniques and interaction design more than any other techniques. This could be recognized to the importance of using objective methods to avoid biased human responses. Furthermore, "machine voice" is a fairly popular usability measure. This could be attributed to the VA developers trying to give the VA a more human and intelligent attributes. The more users perceive the machine voice as intelligent and humanlike, the more they trust and adopt it. More objective technique methods should be created and used on the independent variables when measuring machine voice. Subjective techniques such as Quantitative methods are easy to use and straightforward. However, they can produce biased results.

Interactive design experiments are the most commonly used technique employed to measure the usability. However, the interaction depends on voice modality, making it different from the traditional interaction design that uses visual cues as part of its essential components. Moreover, interaction design also triggers an emotional response, which makes it effective when measuring user attitude. The absence of visual elements in interactive design used might debatably defeat the purpose of clear communication. A new standard of interaction design uniquely for voice modality should be done.

Future Works and Limitation

One limitation in our study was using a few databases as our articles source; in future studies, we intend to add more journal databases such as Scopus, and Taylor and Francis. The majority of the experiment studies we collected was conducted in a controlled environment; future studies will focus on usability measures and independent variables, that are used in natural settings; furthermore, the results can be compared together More studies should be carried out on objective techniques, also how they could cooperate with subjective techniques. This is vital because, with the rise of user expectations of voice assistants, it will be essential to understand how techniques complement each other in each usability measurement.

Conclusion

Our study aimed to understand what is currently employed for measuring voice assistant usability, and we identified the different independent variables, dependent variables, and the techniques used. Furthermore, we also focused on using the ISO 9241-11 framework to measure the usability of voices assistants. Our study classified five independent variable classes used for measuring the dependent variables. These separate classes were categorized based on the similarities between the member groups. Also, our study used the three usability measures in the ISO 9241-11 framework in conjunction with the other three to serve as the dependable variables. We uncovered that voice assistants such as car interface speakers were not studied enough, and currently, smart speakers have the most focus. Dependent variables such as machine voice (anthropomorphism) and attitude recently have more concentration than the old usability measures, such as effectiveness. We also uncovered that usability is dependent on the context of use, such as the same independent variables could be used in different usability measures. Our study highlights the relationship between the independent and dependent variables used by other studies. In conclusion, our study used the ISO 9241-11 to analyse usability. We also highlight what has been carried out on VA's usability and what gaps are left. Moreover, we concluded even though there is a lot of usability measurement carried out, there are still many aspects that have not been researched. Furthermore, the current ISO 9241-11 framework is not suitable for measuring the recent advancement of VA because the user needs and expectation have changed with the rise of technology. Using the ISO 9241-11 framework will create ambiguity in explaining some usability measures such as machine voice, attitude and cognitive load. However, it has the potential to be a foundation for future VA usability frameworks.

Funding This study was funded by The Asahi Glass Foundation.

Declarations

Conflict of Interest The author declares that they have no conflict of interest.

References

1. Hoy MB. Alexa, Siri, Cortana, and more: an introduction to v oice assistants. Med Ref Serv Quart. 2018;37(1):81–8.
2. Zwakman DS, Pal D, Arpnikanondt C. Usability evaluation of artificial intelligence-based voice assistants: the case of amazon Alexa. SN Comput Sci. 2021. https://doi.org/10.1007/s42979-020-00424-4.
3. Segi H, Takou R, Seiyama N, Takagi T, Uematsu Y, Saito H, Ozawa S. An automatic broadcast system for a weather report radio program. IEEE Trans Broadcast. 2013;59(3):548–55.
4. Noel S. Human computer interaction (HCI) based Smart Voice Email (Vmail) Application—Assistant f or Visually Impaired Users (VIU). In: 2020 third international conference on smart systems and inventive technology (ICSSIT) (pp 895–900). IEEE; 2020.
5. Sangle-Ferriere M, Voyer BG. Friend or foe? Chat as a double-edged sword to assist customers. J Serv Theory Pract. 2019;29:438–61.
6. Lugano G. Virtual assistants and self-driving cars. In: 2017 15th International Conf erence on ITS Telecommunications (ITST) (pp 1–5). IEEE; 2017.
7. Rybinski K, Kopciuszewska E. Will artif icial intelligence rev olutionise the student evaluation of teaching? A big data study of 1.6 million student reviews. Assessment & Evaluation in Higher Education; 2020, pp. 1–13
8. Tankovska H. Number of digital voice assistants in use worldwide 2019–2024 (in billions), 2020. https://www.statista.com/statistics/973815/worldwide-digital-voice-assistant-in-use/, (accessed 17 Nov 2021)
9. Pal D, Arpnikanondt C, Funilkul S, Chutimaskul W. The adoption analysis of voice-based smart IoT products. IEEE Internet Things J. 2020;7(11):10852–67.
10. Zwakman DS et al. Voice usability scale: measuring the user experience with voice assistants. In: 2020 IEEE International Symposium on Smart Electronic Sy stems (iSES)(Formerly iNiS). IEEE; 2020.
11. Coronado E, Deuff D, Carreno-Medrano P, Tian L, Kulić D, Sumartojo S, et al. Towards a modul ar and distributed end-user dev elopment framework f or human-robot interaction. IEEE Access. 2021;9:12675–92.
12. Maguire M. Development of a heuristic evaluation tool f or voice user interf aces. In: International conference on human-computer interaction. Cham: Springer; 2019. p. 212–25.
13. Fulfagar L, Gupta A, Mathur A, Shrivastava A. Development and evaluation of usability heuristics for voice user interfaces. In: International conference on research into design. Singapore: Springer; 2021. p. 375–85.
14. Nowacki C, Gordeeva A, Lizé AH. Improving the usability of voice user interfaces: a new set of ergonomic criteria. In: International conference on human-computer interaction. Cham: Springer; 2020. p. 117–33.
15. Pal D, Zhang X, Siyal S. Prohibitive factors to the acceptance of Internet of Things (IoT) technology in society: a smart-home context using a resistive modelling approach. Technol Soc. 2021;66:101683.
16. Murad C, Munteanu C. "I don't know what you're talking about, HALexa" the case f or voice user interface guidelines. In: Proceedings of the 1st International Conference on Conversational User Interfaces,2019; pp. 1–3.
17. Budiu R, Laubheimer P. Intelligent assistants have poor usability: a user study of Alexa, Google assistant, and Siri. Nielsen Norman Group; 2018. Available online at https://www.nngroup.com/articles/intelligentassistant-usability/ (last accessed 4/12/2019).
18. Murphy CN, Yates J. The international organization f or standardization (ISO): global governance through voluntary consensus. Routledge; 2009.
19. ISO 9241-11. Ergonomic requirements for office work with visual display terminals (VDTs)—Part II guidance on usability; 1998.
20. Weichbroth P. Usability attributes revisited: a time-framed knowledge map. In 2018 Federated Conference on Computer Science and Information Systems (FedCSIS) (pp 1005–1008). IEEE; 2018.
21. Petrock V. Voice assistant and smart speaker users 2020. Insider Intelligence; 2020. Retrieved November 22, 2021, from https://www.emarketer.com/content/voice-assistant-and-smart-speaker-users-2020
22. Pinelle D, Wong N and Stach T. Heuristic evaluation f or games: usability principles for video game design. In: Proceedings of SIGCHI Conference on Human Factors in Computing Sy stems (2008); 2008, pp. 1453–1462. https://doi.org/10.1145/1357054.1357282.
23. Sutcliffe A, Gault B. Heuristic evaluation of virtual reality applications. Interact Comput 16. 2004;4:831–49. https://doi.org/10.1016/j.intcom.2004.05.00.
24. Sharif K, Tenbergen B. Smart home voice assistants: a literature survey of user privacy and security vulnerabilities. Complex Syst Inform Model Quart. 2020;24:15–30.
25. de Barcelos Silva A, Gomes MM, da Costa CA, da Rosa Righi R, Barbosa JLV, Pessin G, et al. Intelligent personal assistants: a systematic literature review. Expert Syst Appl. 2020;147: 113193.
26. Gubareva R and Lopes RP. Virtual assistants for learning: a systematic literature review. In: CSEDU (1); 2020, pp. 97–103.
27. Bérubé C, Schachner T, Keller R, Fleisch E, Wangenheim F, Barata F, Kowatsch T. Voice-based conversational agents for the prevention and management of chronic and mental health conditions: systematic literature review. J Med Internet Res. 2021;23(3):e25933.
28. Chi OH, Gursoy D and Chi CG. Tourists' attitudes toward the use of artificially intelligent (AI) devices in tourism service delivery: moderating role of service value seeking. J Travel Res. 2020;0047287520971054.
29. Kim S. Exploring how older adults use a smart speaker-based voice assistant in their first interactions: qualitative study. JMIR Mhealth Uhealth. 2021;9(1): e20427.
30. Coursaris CK, Kim DJ. A meta-analytical review of empirical mobile usability studies. J Usability Stud. 2011;6(3):117–71.
31. Goh ASY, Wong LL, Yap KYL. Evaluation of COVID-19 information provided by digital voice assistants. Int J Digital Health. 2021;1(1):3.
32. Rapp A, Curti L, Boldi A. The human side of human-chatbot interaction: a systematic literature review of ten years of research on text-based chatbots. Int J Hum-Comput Stud. 2021;151: 102630.
33. Seaborn K, Miyake NP, Pennefather P, Otake-Matsuura M. Voice in human-agent interaction: a survey. ACM Comput Surv (CSUR). 2021;54(4):1–43.
34. Castro JW, Ren R, Acuña ST and Lara JD. Usability of chatbots: a systematic mapping study; 2019.

35. Bhirud N, Tataale S, Randive S, Nahar S. A literature review on chatbots in healthcare domain. Int J Sci Technol Res. 2019;8(7):225–31.
36. Ahmad NA, Che MH, Zainal A, Abd Rauf MF, Adnan Z. Review of chatbots design techniques. Int J Comput Appl. 2018;181(8):7–10.
37. Gentner T, Neitzel T, Schulze J and Buettner R. A Systematic literature review of medical chatbot research from a behavior change perspective. In: 2020 IEEE 44th annual computers, software, and applications conference (COMPSAC). IEEE; 2020, pp. 735–740.
38. Cunningham-Nelson S, Boles W, Trouton L and Margerison E. A review of chatbots in education: practical steps forward. In: 30th Annual Conference for the Australasian Association for Engineering Education (AAEE 2019): Educators Becoming Agents of Change: Innovate, Integrate, Motivate. Engineers Australia; 2019, pp. 299–306.
39. Van Pinxteren MM, Pluymaekers M, Lemmink JG. Human-like communication in conversational agents: a literature review and research agenda. J Serv Manag. 2020;31:203–25.
40. Weichbroth P. Usability attributes revisited: a time-framed knowledge map. In: 2018 Federated Conference on Computer Science and Information Systems (FedCSIS). IEEE; 2018, pp. 1005–1008.
41. Bevan N, Carter J, Earthy J, Geis T, Harker S. New ISO standards for usability, usability reports and usability measures. In: International conference on human-computer interaction. Cham: Springer; 2016. p. 268–78.
42. Moumane K, Idri A, Abran A. Usability evaluation of mobile applications using ISO 9241 and ISO 25062 standards. Springerplus. 2016;5(1):1–15.
43. Yahya H, Razali R. A usability-based framework for electronic government systems development. ARPN J Eng Appl Sci. 2015;10(20):9414–23.
44. Alva ME, Ch THS, López B. Comparison of methods and existing tools for the measurement of usability in the web. In: International conference on web engineering. Berlin: Springer; 2003. p. 386–9.
45. He X, Persson H, Östman A. Geoportal usability evaluation. Int J Spatial Data Infrastruct Res. 2012;7:88–106.
46. Dietlein CS, Bock OL. Development of a usability scale based on the three ISO 9241–11 categories "effectiveness, "efficacy" and "satisfaction": a technical note. Accred Qual Assur. 2019;24(3):181–9.
47. Nik Ahmad NA and Hasni NS. ISO 9241–11 and SUS measurement for usability assessment of dropshipping sales management application. In: 2021 10th International Conference on Software and Computer Applications. 2021; pp. 70–74.
48. Kitchenham B. Procedures f or perf orming systematic reviews. Keele Univ. 2004;33(2004):1–26.
49. Seaborn K, Miyake NP, Pennefather P, Otake-Matsuura M. Voice in human–agent interaction: a survey. ACM Comput Surv (CSUR). 2021;54(4):1–43.
50. Al-Qaysi N, Mohamad-Nordin N, Al-Emran M. Employing the technology acceptance model in social media: a systematic review. Educ Inf Technol. 2020;25(6):4961–5002.
51. Kitchenham B and Charters S. Guidelines f or performing systematic literature reviews in software engineering; 2007.
52. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD CD, The PRISMA, et al. statement: an updated guideline f or reporting sy stematic rev iews. BMJ. 2020;2021(372): n71. https://doi.org/10.1136/bmj.n71.
53. Martelaro N, Teevan J and Iqbal ST. An exploration of speech-based productivity support in the car. In: Proceedings of the 2019 CHI conference on human factors in computing systems. 2019; pp. 1–12
54. Jeong Y, Lee J and Kang Y. Exploring effects of conversational fillers on user perception of conversational agents. In: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19). 2019; 1–6. https://doi.org/10.1145/3290607.3312913.
55. Yu Q, Nguyen T, Prakkamakul S and Salehi N. "I almost fell in love with a machine": speaking with computers affects self-disclosure. In: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19). 2019; pp. 1–6. https://doi.org/10.1145/3290607.3312918
56. Kiesel J, Bahrami A, Stein B, Anand A, and Hagen M. Clarifying false memories in voice-based search. In: Proceedings of the 2019 Conference on Human Information Interaction and Retrieval (CHIIR '19). 2019; 331–335. https://doi.org/10.1145/3295750.3298961.
57. Kontogiorgos D, Pereira A, Andersson O, Koivisto M, Rabal EG, Vartiainen V and Gustafson J. The effects of anthropomorphism and non-verbal social behavior in virtual assistants. In: Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents (IVA '19). 2019; 133–140. https://doi.org/10.1145/3308532.3329466
58. Hoegen R, Aneja D, McDuff D and Czerwinski M. An end-to-end conversational style matching agent. In: Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents (IVA '19). 2019; 111–118. https://doi.org/10.1145/3308532.3329473
59. Luo Y, Lee B and Choe EK. TandemTrack: shaping consistent exercise experience by complementing a mobile app with a smart speaker. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20). 2020; 1–13. https://doi.org/10.1145/3313831.3376616
60. Doyle PR, Edwards J, Dumbleton O, Clark L and Cowan BR. Mapping perceptions of humanness in intelligent personal assistant interaction. In: Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '19). 2019. https://doi.org/10.1145/3338286.3340116.
61. Jaber R, McMillan D, Belenguer JS and Brown B. Patterns of gaze in speech agent interaction. In: Proceedings of the 1st International Conference on Conversational User Interfaces - CUI '19 (the 1st International Conference). 2019; 1–10. https://doi.org/10.1145/3342775.3342791.
62. Bortoli M, Furini M, Mirri S, Montangero M and Prandi C. Conversational interfaces for a smart campus: a case study. In: Proceedings of the international conference on advanced visual interfaces (AVI '20). 2020. https://doi.org/10.1145/3399715.3399914.
63. Wu Y, Edwards Y, Cooney O, Bleakley A, Doyle PR, Clark L, Rough D and Cowan BR. Mental workload and language production in non-native speaker IPA interaction. In: Proceedings of the 2nd Conference on Conversational User Interfaces (CUI '20). 2020. https://doi.org/10.1145/3405755.3406118
64. Brüggemeier B, Breiter M, Kurz M and Schiwy J. User experience of Alexa when controlling music: comparison of face and construct validity of four questionnaires. In: Proceedings of the 2nd conference on conversational user interfaces (CUI '20). 2020. https://doi.org/10.1145/3405755.3406122
65. Machine body language: expressing a smart speaker's activity with intelligible physical motion. 57
66. Bartneck C, Kulić D, Croft E, Zoghbi S. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. Int J Soc Robot. 2009;1(1):71–81. https://doi.org/10.1007/s12369-008-0001-3A.
67. Braun M, Mainz A, Chadowitz R, Pfleging B and Alt F. At your service: designing voice assistant personalities to improve automotive user interfaces. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19), 2019;40:1–40:11. https://doi.org/10.1145/3290605.3300270
68. Burbach L, Halbach P, Plettenberg N, Nakayama J, Ziefle M and Valdez AC. "Hey, Siri", "Ok, Google", "Alexa". Acceptance-relevant factors of virtual voice-assistants. In 2019: IEEE

International Professional Communication Conference (ProComm) (ProComm '19), 2019;101–111. https://doi.org/10.1109/ProComm.2019.00025.
69. Pal D, Arpnikanondt C, Funilkul S, and Varadarajan V. User experience with smart voice assistants: The accent perspective. In 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT '19), 2019;1–6. https://doi.org/10.1109/ICCCNT45670.2019.8944754.
70. Chin H, Molefi L, and Yi Y. Empathy is all you need: How a conversational agent should respond to verbal abuse. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20), 2020; 1–13. https://doi.org/10.1145/3313831.3376461.
71. Crowell CR, Villanoy M, Scheutzz M and Schermerhornz P. Gendered voice and robot entities: Perceptions and reactions of male and female subjects. In: Proceedings of the 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2009), 2009; 3735–3741. https://doi.org/10.1109/IROS.2009.5354204
72. Lee S, Cho M and Lee S. What if conversational agents became invisible? Comparing users' mental models according to physical entity of AI speaker. In: Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 2020; 4, 3. https://doi.org/10.1145/3411840
73. Dahlbäck N, Wang QY, Nass C and Alwin J. Similarity is more important than expertise: Accent effects in speech interfaces. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07), 2007; 1553–1556. https://doi.org/10.1145/1240624.1240859
74. Lee EJ, Nass C, and Brave S. Can computer-generated speech have gender?: An experimental test of gender stereotype. In Proceedings of the CHI'00 Extended Abstracts on Human factors in Computing Systems (CHI EA '00), 2000; 289–290. https://doi.org/10.1145/633292.633461
75. Nass C, Jonsson I-M, Harris H, Reaves B, Endo J, Brave S and Takayama L. Improving automotive safety by pairing driver emotion and car voice emotion. In: CHI '05 Extended Abstracts on Human Factors in Computing Systems (CHI EA '05), 2005;1973–6. https://doi.org/10.1145/1056808.1057070.
76. Shi Y, Yan X, Ma X, Lou Y and Cao N. Designing emotional expressions of conversational states for voice assistants: Modality and engagement. In: Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18), 2018;1–6. https://doi.org/10.1145/3170427.3188560.
77. Kim S, Goh J, and Jun S. The use of voice input to induce human communication with banking chatbots. In: Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (HRI Companion '18), 2018;151–152. https://doi.org/10.1145/3173386.3176970.
78. Shamekhi A, Liao QV, Wang D, Bellamy RKE and Erickson T. Face value? Exploring the effects of embodiment for a group facilitation agent. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18),2018;391:1–391:13. https://doi.org/10.1145/3173574.3173965
79. Torre I, Goslin J, White L and Zanatto D. Trust in artificial voices: a "congruency effect" of first impressions and behavioral experience. In Proceedings of the 2018 Technology, Mind, and Society Conference (TechMindSociety '18), 2018. Article No. 40. https://doi.org/10.1145/3183654.3183691.
80. Yarosh S, Thompson S, Watson K, Chase A, Senthilkumar A, Yuan Y and Brush AJB. Children asking questions: Speech interface reformulations and personification preferences. In: Proceedings of the 17th ACM Conference on Interaction Design and Children (IDC '18), 2018;300–12. https://doi.org/10.1145/3202185.3202207.
81. Stucker BE, Wicker R. Direct digital manufacturing of integrated naval systems using ultrasonic consolidation, support material deposition and direct write technologies. UTAH STATE UNIV LOGAN; 2012.
82. Kaplan A, Haenlein M. Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. Bus Horiz. 2019;62(1):15–25.
83. Humphry J and Chesher C. Preparing for smart voice assistants: Cultural histories and media innovations. New media Soc. 2020;1461444820923679
84. Moar JS. Cov id-19 and the Voice Assistants Market. Juniper Research. Retrieved Nov ember 25, 2021, f rom https://www.juniperresearch.com/blog/august-2021/covid-19-and-the-voice-assistants-market
85. Vailshery LS. Topic: Smart speakers. Statista. Retrieved November 25, 2021, from https://www.statista.com/topics/4748/smart-speakers/#:~:text=As%20of%202019%20an%20estimated,increase%20to%20around%2075%20percent
86. Pal D, Vanijja V, Zhang X, Thapliyal H. Exploring the antecedents of consumer electronics IoT devices purchase decision: a mixed methods study. IEEE Trans Consum Electron. 2021;67(4):305–18. https://doi.org/10.1109/TCE.2021.3115847.
87. Pal D, Arpnikanondt C, Razzaque MA, Funilkul S. To trust or not-trust: privacy issues with voice assistants. IT Professional. 2020;22(5):46–53. https://doi.org/10.1109/MITP.2019.2958914.