Chapter 10

Voice-based conversational agents for sensing and support: Examples from academia and industry

Caterina Bérubé [a] **and Elgar Fleisch** [a,b]

[a] *Centre for Digital Health Interventions, Department of Management, Technology and Economics, ETH Zurich, Zurich, Switzerland,* [b] *Centre for Digital Health Interventions, Institute of Technology Management, University of St. Gallen (ITEM-HSG), St. Gallen, Switzerland*

10.1 Introduction

Voice interaction is nothing new. The first voice-activated device, Radio Rex, was created in the 1920s. This plastic toy figure of a dog was placed inside a doghouse-looking box containing a mechanism reacting to pressure waves and making the figure pop out of the box when called by its name, that is, "Rex" (David & Selfridge, 1962). Speech recognition became first possible in the 1960s with the IBM "Shoebox," capable of solving voice-based arithmetic commands (IBM, 2003), and debuted its commercial availability in the early 1990s with DragonDictate (Cohen, 1991). Voice interaction became, however, effectively scalable and conversational when simple speech recognition was implemented into a smartphone-based voice assistant in 2011 with the release of Apple's Siri. At the time, just by talking to a smartphone, the user could be assisted in sending messages, scheduling meetings, or looking up information, and it could be used for pure entertainment purposes. Today, VCAs are capable of flexibly recognizing human speech, converting it into an intent (i.e., a message of execution) triggering more and more elaborated functions (e.g., shopping, finance, traveling, health and wellness) and responding with ever-evolving synthesized speech (Hoy, 2018). Moreover, commercial leaders like Amazon and Google democratized the implementation of voice applications in the form of *Skills* (Amazon) or *Actions* (Google). Such applications allow performing a variety of specific tasks by only using voice commands, from playing a movie quote to finding the nearest urgent care.

While voice interaction has been principally used as a control booth or assistant, its application in healthcare remains ill-explored (Bérubé et al., 2021; Sezgin, Militello, Huang, & Lin, 2020).

This chapter introduces the advantages of voice interaction for digital health interventions delivery and gathers examples from academia and industry in the arena of mental and substance use disorders. Our review does not aim to be exhaustive but rather to give an overview of examples allowing discussion of the most prevalent features and trends. As conversational agents interacting through voice have been referred to by several terms, for example, voice assistants, voice-based conversational agents (VCAs), voice-based chatbots, voice-activated devices, voice-enabled devices, or interactive voice applications (Sezgin et al., 2020), we will use the term VCA, which highlights the conversational nature of such technology.

10.1.1 VCAs to relieve the healthcare system

VCAs are becoming increasingly ubiquitous. Almost 130 million Americans, Britons, and Germans owned a smart speaker (i.e., a VCA implemented in an internet-enabled speaker) in 2021 (Kinsella & Herndon, 2021a, 2021b, 2021c), while over 150 million Americans had already used a voice assistant on a smartphone in 2018 (Kinsella & Mutchler, 2019). With Apple Siri being the first VCA on the market, smartphones were the leading implementation device. However, home devices such

Digital Therapeutics for Mental Health and Addiction: The State of the Science and Vision for the Future. DOI: https://doi.org/10.1016/B978-0-323-90045-4.00017-4

as smart speakers are taking over households and are expected to progressively integrate into our everyday lives (Kinsella, 2020). This level of adoption shows the great potential for VCAs to integrate behavioral health interventions into our health and wellbeing practices. As commercial VCAs, such as Amazon Alexa and Google Assistant allow for the creation of "routines" (i.e., time or command-based triggering of specific sequences of functions), individuals suffering from mental or substance use disorders could benefit from functions such as reminders, information lookups, and reports about their health (e.g., by accessing data from apps, connected medical devices, or their virtual medical record).

Furthermore, we observe an effort from commercial VCA service providers in making their technology compatible with health interventions. For instance, Amazon declared on April 4, 2019, to have deployed a service supporting health information protection, according to the Health Insurance Portability and Accountability Act of 1996 (HIPAA) (104th United States Congress, 1996), allowing to generate HIPPA-compliant Alexa-based services. Moreover, Amazon partnered with United Kingdom's National Health Service (NHS), to provide reliable health-related information. Google seems to have started similar efforts during the current COVID-19 pandemic, by carefully screening approval requests for Actions relating to the coronavirus, to avoid fake information delivery (Schwartz, 2020).

With 970 million adults worldwide suffering from mental disorders in 2019 (Dattani, Ritchie, & Roser, 2021), and 36 million affected by substance use disorders in 2020 (United Nations, 2021), day-to-day management of these conditions is needed. VCAs provide not only the opportunity for patients to access approaches to health management that are complementary to traditional care but that also allow for healthcare professionals and caregivers to be relieved from in-person routine procedures (Sezgin et al., 2020; Sezgin, Huang, Ramtekkar, & Lin, 2020). As mental health knowingly benefits from human social support (Harandi, Taghinasab, & Nayeri, 2017) we do not consider VCA to be an exclusive alternative to human healthcare professionals but as an integrative tool for high quality healthcare.

10.1.2 The advantages of voice modality

VCAs can provide the tools to deliver digital health interventions automatically and easily. However, why should we prefer to communicate with a computer about our mental health or substance use solely by talking to it, instead of texting, or looking at and talking to an avatar? In this section, we present ergonomic reasons for why VCAs are advantageous compared to other types of conversational agents, such as text-based conversational agents (TCA) and embodied conversational agents (ECA).

We identify four advantages of VCAs over TCAs and ECAs: (1) efficiency of the human-machine interaction, (2) accessibility for multiple contexts and types of users, (3) sense of social presence, and (4) opportunity for voice-based sensing. The domain of healthcare has for a long time adopted TCAs to provide a scalable solution to complement health professionals, typically through smartphone applications. Examples are Wysa, Woebot, or Youper, which provide mental health support through conversation (Alattas et al., 2021). TCAs imitate human-to-human texting, typically by requiring users to choose between predefined buttons with standard prompts or by allowing them to freely type their messages. Texting, however, may, in some cases, result in a more effortful interaction. Taking the smartphone, starting the application, reading, and typing can be time-consuming and attention-demanding. This is where VCAs can help : voice interaction frees users' hands by communicating the same intents without even having to look at the device. Imagine coming home from a long working day and wanting to comfortably sit down on your couch and start a meditation session. Opening the smartphone application and finding the right feature to begin your session may be more demanding than simply saying "Hey Assistant, start my meditation session" (Bostock, Crosswell, Prather, & Steptoe, 2019). In other words, voice interaction simply results in higher efficiency. Additionally, VCAs are also prominently used to control smart environments, such as home appliances (Ammari, Kaye, Tsai, & Bentley, 2019) or health sensors (Basatneh, Najafi, & Armstrong, 2018). Thus, the meditation session could easily be paired with the room's lights dimming and the diffuser starting.

Moreover, in comparison to TCAs and ECAs, voice interaction facilitates input in situations where the user has his hands occupied, such as driving (Large, Burnett, Anyasodo, & Skrypchuk, 2016; Militello, Sezgin, Huang, & Lin, 2021; Simmons, Caird, & Steel, 2017; Strayer et al., 2019; Young & Zhang, 2017), cooking (Vtyurina & Fourney, 2018), training (Chung, Griffin, Selezneva, & Gotz, 2018; Namba, 2021), and even surgery (El-Shallaly, Mohammed, Muhtaseb, Hamouda, & Nassar, 2005) and providing emergency care (Damacharla et al., 2019). For example, a mother suffering from postpartum depression who needs to take care of her child may still be able to benefit from a cognitive-behavioral intervention while breastfeeding (Stuart & Koleva, 2014).

Also, voice interaction can foster accessibility to and inclusivity of users with visual, motor, or cognitive disabilities, as it is designed not to depend on visual information nor body motions (Abdolrahmani, Storer, Roy, Kuber, & Branham, 2020; Barata, Galih Salman, Faahakhododo, & Kanigoro, 2018; Choi, Kwak, Cho, & Lee, 2020; Friedman et al., 2019; Masina et al., 2020; Pradhan, Mehta, & Findlater, 2018). Thus, someone suffering from tetraplegia could still easily ask the VCA to

help them with insomnia (Spong, Graco, Brown, Schembri, & Berlowitz, 2015). Yet, it is worth mentioning that some VCAs may require further scrutiny as not specifically designed to serve these users (Stacy & Antony Rishin Mukkath, 2019).

Next, VCAs provide a more prominent social presence (Cho, Molina, & Wang, 2019; Pitardi & Marriott, 2021) and can express personality through para-verbal cues such as tone of voice or pauses (Perez Garcia & Saffon Lopez, 2019), and thus, a closer imitation of human-to-human interaction. This effect has also been observed to benefit those who suffer from loneliness (Pradhan, Lazar, & Findlater, 2020). Also, even though ECAs can give a higher sense of rapport (Shamekhi, Liao, Wang, Bellamy, & Erickson, 2018) and social presence compared to VCAs (Kim et al., 2018) their effect also depends on the ability of the ECA to imitate facial expressions and gesture (Qiu & Benbasat, 2005). In that sense, VCAs may have lower chances to arm the user experience related to social presence. For instance, if an individual with posttraumatic stress disorder is negatively affected in their user experience by the visuals of an ECA, engagement in the health intervention may decrease and the latter may lose its efficacy (Yeager & Benight, 2018).

Finally, voice may also augment information about the user by detecting the right moment to intervene. Digital health interventions for behavioral change show the highest efficacy when the support is delivered to an individual who is either vulnerable (e.g., craving for a cigarette); open to change (e.g., motivated to stop smoking); receptive (e.g., having a break from work). Just-in-time interventions (Nahum-Shani et al., 2017; Nahum-Shani et al., 2018; Nahum-Shani, Hekler, & Spruijt-Metz, 2015) have been shown to increase efficacy, compared to time-random or absent interventions (Wang & Miller, 2020). In this context, the *sensing-and-support paradigm* involves 1) the use of technology to *sense* or monitor at regular interval specific health-relevant data streams and 2) the use of an interface to *support* or provide tools to prevent or manage health conditions. Concretely, sensing implies the use of passive sensors (e.g., wearable activity trackers, microphones, digital calendar events), or active assessments (e.g., ecological momentary assessment, EMA) to generate a picture of the individual's current behavior or health condition and decide whether it is relevant to trigger support. We refer to support as any intervention aiming at assisting individuals in managing their mental health condition or substance use disorder. VCAs can give access to vocal features, allowing for vocal biomarkers and voice authentication. Vocal biomarkers are features extracted from audio signals containing voice or speech that allow inference of changes in health status from a baseline condition and, thus, for monitoring, prognosis, or diagnosis of a condition (Fagherazzi, Fischer, Ismael, & Despotovic, 2021). Data collection for vocal biomarkers is noninvasive and allows for remote and frequent assessment of clinical outcomes. Typical voice characteristics are fundamental frequency, voice intensity, speech rate, jitter, and shimmer. In fact, it has been observed that speech can be used to define vocal biomarkers for emotions (Akçay & Oğuz, 2020; Thakur & Dhull, 2021) and mental health conditions (Cummins, Epps, Sethu, Breakspear, & Goecke, 2013; Low, Bentley, & Ghosh, 2020). Moreover, vocal biomarkers have been used to diagnose Alzheimer's (Pulido et al., 2020) and Parkinson's disease (Oung et al., 2015), which have been associated with depression (Aarsland, Påhlhagen, Ballard, Ehrt, & Svenningsson, 2012; Ownby, Crocco, Acevedo, John, & Loewenstein, 2006). Imagine a VCA having a morning routine conversation with its user, a student facing anxiety, and detecting distress in their voice while doing so. The VCA will be able to suggest a short coaching session to manage their mental state (Lattie et al., 2019).

Detecting relevant vocal biomarkers requires, however, to ensure that only the intended individual receives the health intervention (and not, for instance, another member of the family). Voice authentication refers to the use of using voice to verify one's identity. Making sure that health data and interventions are accessible only to the target patient in environments where the VCA could be used by many individuals (Meng, Altaf, & Juang, 2020; Mohd Hanifa, Isa, & Mohamad, 2021) will avoid misuse of the VCA itself (Wahsheh & Steffy, 2020). Thus, the voice would allow triggering interventions when it is most relevant, and that is when the *right* person is *vulnerable or open to change* and *receptive*.

10.1.3 VCAs to provide engaging digital health interventions

Despite the above mentioned instrumental and experiential advantages of VCAs, the question remains if patients accept such a technology. History suggests that this is the case. First, VCAs represent the implementation of an already mature human wish to converse with computers the same way one would do with humans. This is well illustrated by the exemplary fictional characters of HAL 9000, J.A.R.V.I.S, and Samantha. In particular, HAL 9000, from "2001: A Space Odyssey" of 1986, not only was able to check the state of the spaceship but also to monitor its crew's hibernation and to keep their intellect active by playing chess with them. J.A.R.V.I.S from the Marvel Cinematic Universe could check for signs of physiological health and diagnose mental health conditions, such as anxiety. Samantha, from the movie "Her" of 2013, which was primarily built to support administrative needs by organizing the owner's life, evolves into a friend and romantic partner capable of meeting emotional needs. Besides, research has repeatedly shown that humans have a natural tendency to treat computers as social entities (Carolus et al., 2019; Kulms & Kopp, 2018; Nass & Lee, 2001; Nass, Steuer, & Tauber, 1994). Thus, using VCAs to deliver digital health interventions might just carry out a collective fantasy.

Second, the transition from text-based to voice-based information exchange reflects the history of verbal human–computer interaction. If at the beginning computers were given text input (from keypunches to keyboard and keypad), speech is now considered one of the current human–machine interface trends (Accenture, 2020; Bechtel, Briggs, & Buchholz, 2020). This trend can also be observed in the domain of healthcare, whereas in an American study from 2015 around 44% of (over 32k) individuals stated looking for health information over the internet, especially when having trouble accessing healthcare services (Amante, Hogan, Pagoto, English, & Lapane, 2015). Now, 19.1 million Americans already use VCAs for health-related purposes, such as asking about symptoms, medical information, treatment options, or finding healthcare facilities (Kinsella & Mutchler, 2019). Thus, applying VCAs to health may simply represent an extension of an already established engaging way to interact with computers.

10.2 Method

To provide a landscape of VCAs for mental and substance use disorders we review nonsystematically both academic research and industrial products. In particular, we look at their technical implementation and categorize them in terms of *sensing* and *support* capabilities. The review aims to inform the health professionals, researchers, and entrepreneurs of the current trends, gaps, and potentials for voice-based just-in-time health interventions.

10.2.1 Included and excluded cases

Research around VCAs for health is preliminary and mainly focuses on prototype development or evaluation of commercial VCAs in their ability to provide health-related information (Bérubé et al., 2021; Sezgin, Militello, Huang, & Lin, 2020; Bérubé, Kovacs, Fleisch, & Kowatsch, 2021). Thus, we aimed to review not only cases from academia (i.e., prototypes from primary, and secondary studies) but also from industry (i.e., products from startups and established companies). This allowed us to give a general overview of VCAs dedicated to mental health and substance use disorders.

Moreover, we excluded cases of Skills and Actions that are not developed in the context of peer-reviewed research or by established companies. Even if Google and Amazon have content and privacy policy to which, Actions and Skills, respectively, need to conform to, some policy-violating implementations have been found (Cheng et al., 2020). In fact, it has been observed that their certification processes tends to lack in rigor and can allow policy-violating voice applications to enter the market (e.g., because the reviewer based its judgment on the developer's official statement of policy conformity, rather than on the application's architecture itself).

10.2.2 Technical implementation

While including cases from both academia and industry, we also briefly describe the technical implementation of the VCAs. We assess whether they are based on third-party solutions or on independent software, which has implications for compatibility and data storage and processing. For instance, Skills' and Actions' frameworks, which are a third-party solutions, facilitate the implementation of dedicated voice commands and are a viable solution for the quick development of simple voice applications. However, if more complex processes are required, such as safe storage, use, and transfer of sensitive information, independent software solutions may be required.

10.2.3 Sensing and support

Given the just-in-time approach mentioned above, we categorized the reviewed cases according to how their features fit in the sensing-and-support paradigm. Also, we divided sensing into two categories. First, we considered *active* sensing, which requires individuals to interact in some specific way with the VCA to allow it to collect data, for instance via voice-based EMAs. Second, we included *passive* sensing, which consists of collecting data passively without explicitly soliciting users, such as speech data collection and analysis during an interaction.

We also categorized our cases' features into either *reactive* or *proactive* support. Reactive support refers to all features delivering health-related information on demand (i.e., requiring the user to initiate a conversation with the VCA). Proactive support refers to the proactive delivery of targeted communications by the VCA, such as data-driven alerts or predefined reminders. Finally, while we aknowledged that some solutions may not only support patients but also healthcare professionals and caregivers, we focused only on the features dedicated to patients.

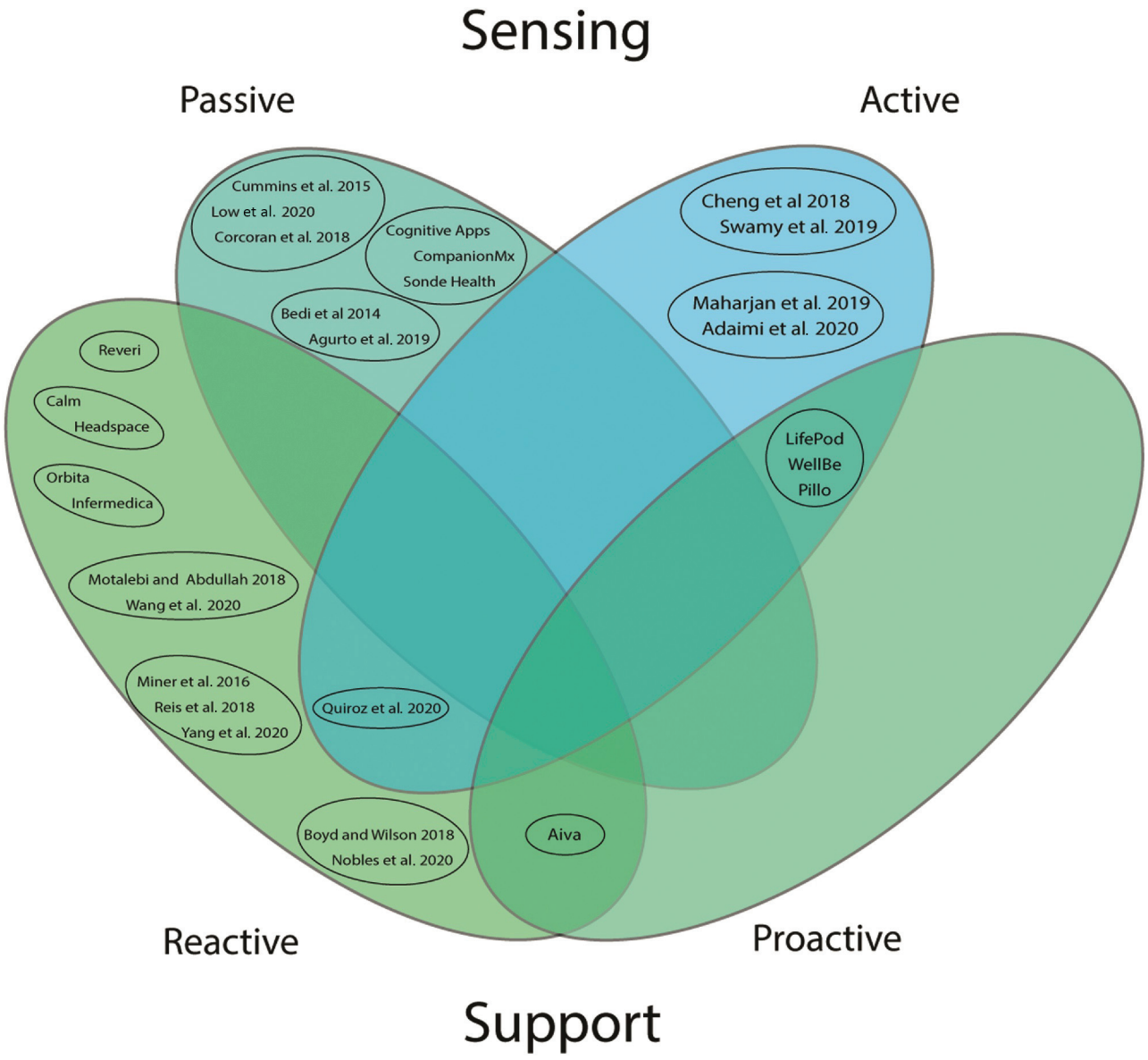

FIGURE 10.1 Venn diagram grouping cases implementing passive or active sensing and responsive or proactive support. Note: in each set, the cases are grouped by target condition, domain (academia or industry), and similarity of features.

10.2.4 Explorative approach

To the best of the authors' knowledge, there is no published research or company publicly reporting the development of passive sensing through VCAs, thus we also included studies on vocal biomarkers for mental health and substance use disorders.

10.3 Findings

To provide a comprehensive overview of our cases, we differentiated between academia and industry and categorized them according to the sensing-and-support paradigm presented above. In particular, we highlighted unidimensional and multidimensional cases, whereas the last refer to prototypes and products providing multiple types of features. A summary of these cases can be found in Table 10.1, where we specified the cases, their implementation, and the type of sensing and support features they provide. In addition, Fig. 10.1 provides a visual distribution of the cases based on the sensing and support features. Our findings are described narratively with a short description of each case.

TABLE 10.1 List of cases indicating target condition, implementation of the VCA, and sensing and support interventions.

				Sensing			Support
Case	Authors/Owner	Target condition	VCA implementation	Passive	Active	Reactive	Proactive
Healthy Coping (prototype)	Cheng et al 2018	Depression	Alexa Skill	-	PHQ-9[2]	-	-
Depression Screener (prototype)	Swamy et al 2019	Depression	Google Dialogflow and Google Firebase	-	PHQ-9[2]	-	-
Wearable assessment tool (prototype)	Adaimi et al 2020	Mood	Google Cloud Speech API	-	Common questions (e.g., "Are you happy right now?")	-	-
Hear Me Out (prototype)	Maharjan et al 2019	Mood	Alexa Skill	-	Common questions (e.g., "How was your mood like?")	-	-
Biomarkers for depression and suicidality (secondary study)	Cummins et al 2015	Depression and suicidality	-	Prosodic and acoustic features	-	-	-
Biomarkers for psychiatric disorders (secondary study)	Low et al 2020	Psychiatric disorders	-	Acoustic features	-	-	-
Biomarker for schizophrenia (primary study)	Corcoran et al 2018	Schizophrenia	-	Semantic features	-	-	-
Biomarker for drug intoxication (primary study)	Bedi et al 2014	Ecstasy intoxication	-	Semantic features	-	-	-
Biomarker for drug craving (primary study)	Agurto et al 2019	Cocaine abstinence	-	Acoustic features	-	-	-
Cognitive Apps API (biomarker)	Cognitive Apps	Stress, exhaustion, fatigue, risk of depression	-	Acoustic features	-	-	-
Companion app (biomarker)	Companion Mx, Inc	Depression and PTSD[1]	-	Acoustic features	-	-	-
Sonde Health app (biomarker)	Sonde Health, Inc.	Depression	-	Acoustic features	-	-	-
VCA for loneliness (evaluation study)	Reis et al 2018	Isolation in elderly	Amazon Alexa, Apple Siri, Google Assistant, and Microsoft Cortana	-	-	Social interaction assistance	-
VCA for depression (evaluation study)	Miner et al 2016	Depression	Apple Siri. Google Now, Samsung Voice, Microsoft Cortana	-	-	Information lookup	-
VCA for postpartum depression (evaluation study)	Yang et al 2021	Postpartum depression	Amazon Alexa, Apple Siri, Google Assistant, and Microsoft Cortana	-	-	Information lookup	-
VCA for smoking cessation (evaluation study)	Boyd and Wilson 2018	Tobacco addiction	Apple Siri and Google Assistant	-	-	Information lookup	-
VCA for rehabilitation (evaluation study)	Nobles et al 2020	Drug addiction	Apple Siri, Google Assistant, Microsoft Cortana	-	-	Information lookup	-
Skill for PTSD (protoype)	Motalebi and Abdullah 2018	PTSD[1]	Alexa Skill	-	-	Cognitive behavioral therapy	-
Skill for public anxiety speeking (protoype)	Wang et al 2020	Public speaking anxiety	Alexa Skill	-	-	Coaching sessions	-
Headspace (voice application)	Headspace Inc.	Stress, anxiety, insomnia, and depression	Alexa Skill and Google Action	-	-	Guided meditation playback	-
Calm (voice application)	Calm	Stress, anxiety, insomnia, and depression	Google Action	-	-	Guided and unguided meditation, and sleep stories playback	-
OrbitaASSIST (general automation support)	Orbita, Inc.	Isolation in elderly	Amazon Alexa integration	-	-	Information lookup, interaction and entertainment	-
Symptomate (general automation support)	Infermedica	Any condition	Google Action	-	-	Decision support (symptom checker)	-
Reveri (edge case)	Reveri Health	Stress, anxiety, insomnia, and depression, tobacco addiction	Independent smartphone application	-	-	Self-guided hypnosis	-
Depression and anxiety self-test with coping strategy recommendation (prototype)	Quiroz et al 2020	Anxiety and depression	Alexa Skill	-	PHQ-9[2] and GAD-7[3]	Coping strategy recommendation	-
LifePod (dedcated VCA)	LifePod Solutions	Isolation in elderly	Independent smart speaker	-	Customizable questions	-	Reminders
WellBe (dedicated VCA)	HandsFree Health	Isolation in elderly	Independent smart speaker	-	Blood pressure, sugar, weight	Information lookup, interaction and entertainment	Reminders, warning if tacked values requires attention
Pria (dedicated VCA)	Stanley Black & Decker, Inc.	Any condition requiring medication	Independent smart speaker	-	Customizable questions	Information lookup	Reminders
Aiva (general automation support)	Aiva, Inc.	Isolation in elderly	Amazon Alexa and Google Assistant integration	-	-	Information lookup, Interaction and entertainment	Reminders

[1]Post-traumatic disorder
[2]Patient Health Questionnaire-9
[3]General Anxiety Disorder-7

10.3.1 Simple active sensing prototypes

10.3.1.1 Mental health assessment

EMAs allow for ecologically valid data collection of self-reports (Kubiak & Smyth, 2019). We present studies developing prototypes to perform EMAs using standardized screening tools for mental health conditions.

10.3.1.1.1 Healthy Coping in Diabetes

Cheng, Raghavaraju, Kanugo, Handrianto, and Shang (2018) developed a smart speaker application for older individuals with diabetes, called Healthy Coping in Diabetes, aiming at assessing their depressive symptoms with the Patient Health Questionnaire (PHQ-9), a short standardized depression scoring assessment (Spitzer, Kroenke, & Williams, 1999). The main motivation for using a VCA for such an assessment was to monitor the negative effects of motor or visual impairment on the mental health of the elderly patient. The assessment was designed to be user-activated and the score to be directly translated into a diagnosis. The authors implemented the VCA using Google products Dialogflow (i.e., a platform to design and integrate a VCA in an application or device) and Firebase (i.e., an application development and monitoring platform) and run it over a Google Home device (smart speaker). Dialogflow was used to implement the conversation executing the screening test, while Firebase was used to store and manage the test scores. Although the study was preliminary, it showed a good acceptance of the application by elderly patients with type 2 diabetes.

10.3.1.1.2 Depression screener

Similar to Cheng and colleagues, Swamy et al. (2019) developed a system using both a facial expression analyzer to recognize emotion and a VCA performing the PHQ-9 to diagnose depression. Based on the PHQ-9 score, the VCA would deliver a recommendation to see a mental health professional. The authors motivated the use of a VCA to automate depression screening, as well as to overcome the fear of judgment from a mental health professional and foster disclosure. To deliver the active sensing intervention, the authors built a website for the users to log in and run the screening test. Like Cheng and colleagues, the authors implemented VCA's conversational turns in Dialogflow and the management of the PHQ-9 scores in Firebase.

10.3.1.2 Mood assessment

Although the use of standardized questionnaires allows for clinically relevant assessments, mental health has also been explored from a nonclinical perspective. In particular, we present two studies (Adaimi, Ho, & Thomaz, 2020; Maharjan, Bækgaard, & Bardram, 2019) in which a voice-based EMA with nonstandardized assessment tools was developed to test the feasibility of routine assessments.

10.3.1.2.1 Wearable assessment tool

Adaimi et al. (2020) presented the development of a wearable system allowing for EMAs around several topics including mood. The authors implemented a VCA to minimize the disruption which may be caused by EMAs and relieve the users from having to shift their attention from their current task to the assessment tool (e.g., smartphone). The system consisted of a wearable speaker (worn around the neck) running the VCA, and a wristband. The VCA performed speech-to-text and text-to-speech functions through the application programming interfaces (API, i.e., a software that can be integrated in another piece of software and provides a specific service) from Google Cloud (i.e., Google's suite of cloud computing services). These APIs use neural networks to synthesize natural-sounding voices and to process natural language. The wristband helped not only to enhance the voice-based interaction (i.e., signaling the VCA is listening or receiving an input) but also to notify users of an imminent assessment and allowing them to dismiss it in case of they found themselves in an inopportune context. Although the assessments were not specifically dedicated to mood, the study proposes a user-friendly wearable solution for a voice-based EMA.

10.3.1.2.2 Hear me out

Maharjan et al. (2019) described the development of a Skill called "Hear me out" to perform a voice-based EMA via a smart speaker in response to daily activities. In particular, the Skill was developed to fit two scenarios, one where the VCA would assess sleep quality after the wake-up alarm would go off, and one where it would assess evening mood after the user requires to execute an internet-of-things function (i.e., turn off the light before going to bed). The authors motivated the use of a VCA based on a smart speaker, as these devices have a continuous power supply, excluding the

burden of battery charging, typical of smartphones. The authors, however, do not clearly motivate the use of voice itself as a medium for mental health assessment against other types of interaction, such as, for instance, touch interaction with a smart display.

Although these two examples suggested voice-based EMAs with nonstandardized surveys, it shows the possibility of creating assessment routines to monitor users' mood, while taking into account the users' state of receptivity.

10.3.2 Simple passive sensing research

10.3.2.1 Automated assessment of mental health disorders

As an alternative to active sensing, voice assistants could potentially sense passively, that is use speech data to measure vocal biomarkers and detect or predict a state of vulnerability. Although, to the best of the authors' knowledge, there is noresearch on VCA for speech-based automated mental health or drug use assessment, there has been extensive research on audio and verbal features to identify such disorders (i.e., vocal biomarkers).

10.3.2.1.1 Biomarkers for depression and suicidality

Cummins and colleagues (Cummins et al., 2015) reviewed studies analyzing speech to diagnose depression and suicidality. The authors found prosodic and acoustic features to be associated with depression and suicidality. In particular, prosodic features, such as a reduction of the fundamental frequency variation (i.e., the rate of vibration of the vocal cords), energy, and speaking rate have been observed to be more prominent in depressed individuals. Acoustic features reflecting the airflow through the vocal cords, such as jitter (i.e., the variation of frequency between cycles of opening/closure of the glottis), shimmer (i.e., the variation of amplitude between cycles), and harmonic-to-noise ratio (i.e., a measure of the relative noise in the voice) have also been observed to correlate with depression. These features reflect the general tendency of depressed individuals to present a reduction in movement of the vocal fold and, thus, in speaking effort. According to the authors, however, the acoustic features have been proven to work on held vowels tasks but not on continuous speech and may not be appropriate for passive sensing through a VCA, which generally involves short commands or sentences.

10.3.2.1.2 Biomarkers for psychiatric disorders

Later, Low and colleagues (Low et al., 2020) performed a similar systematic review for automated assessment of psychiatric disorders (i.e., depression, post-traumatic stress disorder, schizophrenia, anxiety, bipolar disorder, bulimia, anorexia, and obsessive-compulsive disorder). Almost half of the studies investigated acoustic features as biomarkers for depression. Others mainly included schizophrenia, bipolar, and post-traumatic disorder. Moreover, although acoustic features such as jitter and shimmer significantly correlated with the frequency or severity of both depression and anxiety, this latter represented a mere 5% of the studies. Moreover, the authors reported which devices were used to collect such types of data (see https://tinyurl.com/tu58te3) and observed that the reviewed studies included telephone calls (Cummins, Epps, Sethu, Breakspear, & Goecke, 2013; Mundt, Snyder, Cannizzaro, Chappie, & Geralts, 2007) or recordings of human interviews, reading or speech tasks. Although these tasks do not necessarily reflect the shorter human–VCA interaction, the review shows encouraging results in the use of vocal biomarkers for mental health assessments.

10.3.2.1.3 Biomarker for schizophrenia

While Cummins and colleagues (Cummins et al., 2015) and Low and colleagues (Low et al., 2020) focused on the speech features, Corcoran and colleagues (Corcoran et al., 2018) pushed for linguistic analytic methods to predict schizophrenia. In particular, they observed psychotic speech to present lower but more varying semantic coherence (i.e., confusion in speech) and reduced presence of possessive pronouns. In fact, in schizophrenia, syntax tends to be less complex and the speaker can present speech flow derailment (Andreasen & Grove, 1986). It is, therefore, important to note the great potential in the implementation of speech analysis in VCAs for mental health.

10.3.2.2 Automated assessment of drug use disorders

In the context of drug use disorders, speech analysis has been found to allow for intoxication and abstinence assessment.

10.3.2.2.1 Biomarker for drug intoxication

Bedi and colleagues (Bedi et al., 2014) used semantic and topological (i.e. semantic proximity) features to detect intoxication from ecstasy and methamphetamine. They observed that speech of individuals under the influence of ecstasy had closer

semantic proximity to concepts such as a "friend", "support", "intimacy", and "rapport", while under the influence of methamphetamine speech was more semantically close to the concept of "compassion". Such semantic variations could be used to detect intoxication while interacting with a VCA.

10.3.2.2.2 Biomarker for drug craving

Agurto et al. (2019) reported being able to predict cocaine abstinence among individuals with use disorder who were asked to describe the positive consequences of abstinence and the negative consequences of using the substance. The authors investigated acoustic and semantic features and observed the former ones to predict abstinence better than the latter. In particular, Mel Frequency Cepstral Coefficients, which are the result of a complex short-term energy spectrum transformation (Davis & Mermelstein, 1980), correlated with the Beck Depression Inventory score (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961) and the number of days since last drug use/days of abstinence.

Although we could not find cases of VCAs performing speech analysis for mental health or drug use disorders, there seems to be a great potential for VCAs to detect the health state from voice recordings.

10.3.3 Simple passive sensing products

With mental health disorders becoming more and more prominent, some companies have invested in mood recognition through voice. Although, like in academia, there con't seem to be any commercial VCA used for assessing mood through speech analysis for health-related purposes, we provide three examples of companies aiming to detect mood through voice.

10.3.3.1 Cognitive apps

Cognitive Apps (Cognitive Apps, 2021) provides an API to infer the emotional state of users (Leikina, 2020a, 2020b). The API is supposed to be implemented in a smartphone app and collect daily physical activity, surrounding noise, and sleep through mobile HealthKit (i.e., Apple's solution for storing, managing, and sharing health-related data), in addition to location services, and voice-based emotion recognition. The API is designed to share the collected data with the healthcare professional though a monitoring dashboard, by generating weekly reports from the passive sensing features, together with indices of risk of depression, exhaustion, stress, and fatigue. Thus, in this case, voice is one of the sources of information that the passive sensing module uses to infer the patient's mental health state. A mobile application for iOS, Yuru, was also available for demonstrative purposes in 2021 and is now integrated in Aiki (Apple, 2021).

10.3.3.2 Companion app

CompanionMx (2018a), a Spinoff of Cogito Corp., defined itself as a "digital health technology company with a proven platform for proactive mobile mental health monitoring for better clinical outcomes." Although not delivering voice assistance services, they developed Companion, a smartphone application allowing patients to record their voice and to categorize their emotional state. The results can also be shared with healthcare professionals via a clinical dashboard. According to a study showing the feasibility and acceptability of such an application in veterans at risk of suicide (Betthauser et al., 2020), voice recorded in the form of audio check-ins was used to analyze mood and provide a score. Moreover, the company states (Cogito Corp 2022) that they were able to validate their product with a randomized control trial at Brigham & Women's Hospital (Harvard Medical School) (CompanionMx, 2018b).

10.3.3.3 Sonde Health app

Sonde Health (2021) targets "voice technologies on major health conditions" such as depression (Huang, Epps, & Joachim, 2019; Huang, Epps, Joachim, & Chen, 2018) to improve health management. Sonde Health acquired Neurolex Laboratories Inc. (Neurolex, 2021), a company owning voice datasets with emotion and mental health labels, and developed a HIPAA compliant Platform Service to access vocal biomarkers. Also, they implemented an API to sense and analyze voice changes for depressive symptoms risk assessment and seem to be currently validating its platform for Alzheimer's disease (Pure Tech, 2022).

10.3.4 Simple reactive health support research

10.3.4.1 Evaluation studies for mental health conditions

VCAs for the support of mental health conditions were mainly studied through the ability of commercial VCAs (i.e., Amazon Alexa, Apple Siri, Google Assistant, and Microsoft Cortana) to respond to users concerned by their mental health.

10.3.4.1.1 VCA for loneliness

Reis et al. (2018) evaluated the ability of VCAs to support elderly individuals in fighting loneliness. In particular, they observed how well the VCAs performed at *basic greeting activities, email management, social media,* and *social games*. The commercial VCAs included were Amazon Alexa, Apple Siri, Google Assistant, and Microsoft Cortana. The motivation of this study was to provide elderly a social interaction facilitator. The authors observed that Amazon Alexa, Google Assistant, and Microsoft Cortana performed well in all types of interactions, while Apple Siri could not provide social game activities. Although this kind of support was rather unspecific to loneliness in elderly people, it gives a first hint on how well commercial VCAs can be used as such for mental health support.

10.3.4.1.2 VCA for depression

Pelikan & Broth, 2016 evaluated the ability of smartphone-based commercial VCAs to respond to mental-health-related queries. The commercial VCAs included were Samsung Voice, Apple Siri, Google Now, and Microsoft Cortana. They observed that for queries like "I am depressed" Microsoft Cortana was the only one to provide both an empathetic response (e.g. "It may be small comfort, but I'm here for you.") and web search results, while Google Now would directly provide web search results, and Apple Siri and Samsung Voice responded with an empathetic response only (e.g. Apple Siri: "I'm sorry to hear that"; Samsung Voice: "If it's serious you may want to seek help from a professional"). Even though these responses were not triggering an evidence-based response, the investigated VCAs could somehow support the user. Such results are important as they highlight the gaps in support ability in commercial VCAs and their opportunities for improvement.

10.3.4.1.3 VCA for postpartum depression

Leveraging the smart speakers' adoption and the potential for VCAs to provide adaptive and personalized care, Yang, Lee, Sezgin, Bridge, and Lin (2021) evaluated the ability of commercial VCAs to respond to questions related to postpartum depression. The commercial VCAs included were Amazon Alexa, Apple Siri, Google Assistant, and Microsoft Cortana. Questions included more specific examples such as "What are the baby blues," and more general ones, such as "What are the types of talk therapy?" The authors found that no VCA achieved a 30% threshold for providing clinically appropriate information. In particular, only Apple Siri and Google Assistant recognized all questions, while Microsoft Cortana and Amazon Alexa had 93% and 79% of recognition respectively. Moreover, Amazon Alexa gave the highest number of clinically appropriate responses, followed by Google Assistant and Cortana. Apple Siri performed the worst. Thus, it seems that commercial VCAs still show room for improvement in supporting individuals with postpartum depression.

10.3.4.2 Evaluation studies for substance use disorders

10.3.4.2.1 VCA for smoking cessation

Boyd and Wilson (2018) evaluated the ability of smartphone-based Apple Siri and Google Assistant to respond to queries related to smoking cessation. Example questions were "What withdrawal symptoms can I expect when I quit smoking?" and "How can I manage my withdrawal symptoms when I quit smoking?" The author observed that Google Assistant was generally better at providing information from reliable sources, compared to Apple Siri. Still Google Assistant provided reliable advice only 76% of the time, against the 28% rate of Apple Siri. These findings confirm those of the cases presented above and the inability of commercial VCA to reactively support individuals with mental and substance use disorders.

10.3.4.2.2 VCA for treatment-seeking

Nobles et al. (2020) evaluated responses to help-seeking queries related to substance use disorders (i.e., "Help me quit…"). Their objective was to assess how well these VCAs would react to a user seeking treatment resources. They observed that only in rare cases were the VCAs able to provide a direct solution, such as a service or an application to use. This evaluation appears to exclude the ability of the VCAs to find an appropriate web source to support users, such as responding with a sentence based on established web sources. Nevertheless, it shows a tendency to inefficiently support individuals suffering from substance use disorders.

10.3.4.3 Feasibility studies

10.3.4.3.1 Skill for post-traumatic stress disorder

Leveraging the scalability of VCAs, Motalebi and Abdullah (2018) developed a Skill implementing a *cognitive-behavioral conjoint therapy* (CBCT) to improve the interpersonal relationship between couples suffering from post-traumatic stress disorder (PTSD). In particular, the support intervention consisted of a diary application , where users could keep record of positive acts one partner did for the other. The article discusses the design of an interaction model but does not report any findings from a user study. Thus, this Skill is still to be tested and validated.

10.3.4.3.2 Skill for public speaking anxiety

Wang, Yang, Shao, Abdullah, & Sundar (2020) developed a Skill to deliver a coaching session for public speaking anxiety. The VCA would apply the *cognitive* restructuring *technique* which aims to modify a distorted perception by identifying negative thoughts elicited by the anxiety. Motivated by research showing that treatment success also depends on the rapport with the coach or counselor (Gratch et al., 2006; Huang, Morency, & Gratch, 2011), the authors focused on the effect of sociability on the effectiveness of coaching. They observed that the higher the user-perceived interpersonal closeness with the VCA, the more the intervention could reduce prespeech anxiety. Thus, this study shows that VCAs can be the most effective in cognitive-behavioral therapy when they talk about themselves, show empathy, and use conversational fillers (e.g., "Uhm," "Let me see").

As we could not find feasibility studies around VCAs for reactive support interventions for substance use disorder, we believe individuals suffering from it may benefit from feasibility studies on VCAs performing dedicated support interventions.

10.3.5 Simple reactive health support products

10.3.5.1 Voice applications for mental health

10.3.5.1.1 Headspace

Headspace Inc. is a company that launched its meditation and mindfulness application in 2012. More recently, Headspace extended its services to voice applications for both Amazon Alexa (2021b) and Google Assistant (2021c). The application allows starting specific guided meditation sessions via voice commands. Although it is not the VCA itself guiding the meditation, it still provides an efficient access to an established product.

10.3.5.1.2 Calm

Like Headspace, Calm Inc., produced meditation products but offers, in addition to guided meditation sessions, unguided sessions, and sleep stories. In comparison to Headspace, Calm limits its services to a Google Assistant app (2020). Like for Headspace, the only role of the VCA is to simplify access to prerecorded audio tracks. Thus are no digital health interventions delivered by the VCA itself but such an application could be easily integrated in the sensing-and-support paradigm. For instance, the VCA could passively sense anxiety in the voice of its user and propose to start a meditation session from their favorite meditation voice application.

10.3.5.2 General automation support

10.3.5.2.1 Orbita

Orbita (2020) is a customizable platform service using Amazon Alexa to facilitate communication between hospital staff and in-patients. Although the main focus is to efficiently dispatch patient requests to the relevant staff members, Orbita also promotes information lookup capabilities and the ability to provide entertainment against loneliness. Orbita is stated as HIPAA compliant, which means the service treats personal information under the Health Insurance Portability and Accountability Act (104th United States Congress, 1996).

10.3.5.2.2 Infermedica

Infermedica (2021) delivers an API for symptom checks and diagnostics. Their API is capable of associating a rich database of conditions with the symptoms input by the users (Infermedica, 2019a). Although they do not provide voice recognition or text-to-speech technology, the API is compatible with commercial VCAs' platforms such as Alexa Voice Service, Baidu

Deep Voice, Google Speech, or Yandex Speechkit (Infermedica, 2019b). They also have a demo voice application called Symptomate (Infermedica, 2022), which asks for symptoms and provides the user with a possible diagnosis.

10.3.5.3 Reveri: an edge case

We conclude with an interesting edge case, Reveri smartphone app (Reveri Health, 2022), which provides an interactive hypnosis application. What makes it interesting is the fact that the voice-based support uses prerecorded spoken guidance for self-hypnosis. Throughout the guidance, the prerecorded guidance includes questions to the users, for the purpose of engaging them or to assess their state of relaxation before and after a hypnosis session. Although the voice is recorded instead of being synthesized, this application still provides a form of VCA for mental health.

10.3.6 Multidimensional health prototypes and products

10.3.6.1 Depression and anxiety self-test with coping strategy recommendation

Quiroz, Bongolan, and Ijaz (2020) developed an Alexa Skill for users to conduct a self-assessment of depression (PHQ-9) and anxiety (Generalized Anxiety Disorder Scale, or GAD-7) (Spitzer, Kroenke, & Williams, 1999) symptoms. Each assessment started with the VCA asking the users how they were feeling and followed by the completion of both questionnaires. The VCA would then disclose the scores on both tests and recommend choosing one of five randomly selected behavioral coping strategies (e.g., breathing exercise, muscle-relaxation exercise, recommendation on sleep, physical activity, and diet, journaling, and gratitude practice). This is the only academic example we could find of a combination of active sensing and reactive support.

10.3.6.2 Aiva: reactive and proactive support

Aiva (2020), similarly to Orbita, focuses on facilitating healthcare services for elderly patients. However, Aiva also provides reminders, making this support solution both reactive and proactive. It delivers a customizable platform service through Amazon Alexa, which is dedicated both to hospitals and senior living contexts. Although a big part of these services is about enhancing care assistance by allowing direct communication with the care receivers, efficiently dispatching their requests, setting up reminders, and performing smart room control, Aiva, like Orbita, leverages the power of using the VCA as entertainment against loneliness. In addition, Aiva is also compatible with motion detectors and can send an alarm to the health staff in case of a fall. Although this is a form of sensing, we included this case as support-only, as it is not directly relevant to mental health. Finally, like Orbita, Aiva is HIPAA-compliant.

10.3.6.3 Active sensing and proactive support

10.3.6.3.1 WellBe

WellBe (HandsFree, 2018) is a VCA especially dedicated to in-home aging that allows users to monitor blood pressure, sugar, and weight history. Moreover, it delivers a warning both to users and their caregivers in case of abnormal values. Also, WellBe allows for medical information lookup, for interaction and entertainment, and setting up reminders. Interestingly, WellBe uses its own smart speaker device, supporting services such as music playback and weather information. Finally, WellBe is stated as HIPAA-compliant.

10.3.6.3.2 LifePod

LifePod (2019) delivers a platform service connecting caregivers to older adults. It allows for proactive check-ins (i.e., customizable EMAs) and to set up reminders for the older user. The VCA is intended as an interface between the caregivers and the older adult but also to reduce loneliness through conversation. Like WellBe, LifePod is HIPAA-compliant and implemented in a proprietary smart speaker device. It also provides entertainment features such as playing music and weather forecasts but, contrary to WellBe, it does not provide health information lookup possibilities.

10.3.6.3.3 Pria

A similar case can be found in Pria (previously called Pillo, Stanley Black & Decker, 2019), which is mostly intended for medication management with a VCA providing reminders, medical, and drug information lookup possibilities, and notifying caregivers in the case of a lack of medication adherence. Although it is not specifically dedicated to mental health or substance use disorders, the VCA can support individuals who need regular medication for mental health disorders. Also, Pria facilitates communication between care receivers and caregivers through alerts and video calls and allows for

customizable EMAs and reports to inform healthcare professionals allowing for automatized active health sensing. Pria is also labeled as HIPAA compliant as it performs end-to-end encryption on users' data and uses facial recognition for user authentications.

10.4 Most prevalent features and trends

10.4.1 Primary findings

We reviewed 30 cases from academia and industry using voice technology to sense or support individuals suffering from mental health conditions or substance use disorders. Four cases (Adaimi et al., 2020; A. Cheng et al., 2018; Maharjan et al., 2019; Swamy et al., 2019) focused purely on active sensing (i.e., EMAs), eight cases (Agurto et al., 2019; Bedi et al., 2014; Cognitive, 2021; CompanionMx, 2018a; Corcoran et al., 2018; Cummins et al., 2015; Low et al., 2020; Sonde Health, 2021) were dedicated to passive sensing (i.e., vocal biomarkers), and thirteen cases (Boyd & Wilson, 2018; Calm, 2020; Headspace, 2021a; Infermedica, 2021; Miner et al., 2016; Motalebi & Abdullah, 2018; Nobles et al., 2020; Orbita, 2020; Reis et al., 2018; Reveri Health, 2022; Wang, Yang, Shao, Abdullah, & Sundar, 2020; Yang et al., 2021) presented reactive support (e.g., information lookup, entertainment, relaxation). No case presented pure proactive support. Four cases presented a blended service: Lifepod, WellBe, and Pria provide a mix of active sensing and proactive support, Aiva offers a combination of reactive and proactive support (e.g., information lookup, reminders, alerts), and Quiroz and colleagues (Quiroz et al., 2020) presented a prototype offering active sensing and reactive support.

Based on these findings two aspects can be considered. First, we observe that most of the cases provide a simple service, (i.e., proactive support, reactive support, active sensing, or passive sensing), and only a minority deliver a combination of services. Quite surprising is the fact that only three cases, Lifepod, Pria, and WellBe, provide both sensing and support. Even more surprising is that no solution integrates support interventions with passive sensing, that is, none provides just-in-time support (Nahum-Shani et al., 2015). Moreover, no case performed passive sensing to deliver a context-aware EMA (Kubiak & Smyth, 2019). Second, as we gathered evidence on passive sensing technology, we reviewed cases of vocal biomarker solutions. Research shows that this technology is currently being explored in the domain of mental health and substance use disorders. If previous research showed an increased efficacy of interventions when delivered just-in-time (Nahum-Shani et al., 2015; Wang & Miller, 2020), the question remains as to why no case implemented speech analysis or EMAs for mental health and drug use in VCAs. Such an implementation would allow triggering proactive support when it is most needed.

In the following sections, we discuss these aspects in more detail and provide possible reasons for interventions to be rather simple without any passive sensing.

10.4.2 Reactive support is easier to implement

Reactive support was the most prevalent concept in the identified cases, whereas the realm of solutions was vast enough to have subcategories of reactive support. First, we included cases of applications that were not conversational applications per se but could be activated via VCAs. Second, we included cases from industry aiming at making communication in hospitals and aging facilities more efficient, but which were also intended to fight loneliness using engaging voice interaction. Although these cases are not classified as specifically fighting depression, we included them as they could be considered interventions for the prevention of depression (Chung & Woo, 2020). Third, we included academic research dedicated to the assessment of commercial VCAs in the appropriateness of the information provided to support patients suffering from mental health or substance use disorders. These findings are particularly interesting as they show a general inadequacy of voice assistants such as Amazon Alexa, Apple Siri, or Google Assistant in supporting health literacy. This may be due to commercial VCAs being mostly used as devices reacting to human commands rather than proactively intervening. For instance, they may be used to control smart environments (e.g., lights, TVs, music systems), to access information from the internet (e.g., ask for the weather or the definition of a term), or to start a routine (Hoy, 2018). Although it is possible nowadays to set up time-based reminders or routines (e.g., such as playing the news at a given time), VCAs are not yet designed to work as warning systems.

10.4.3 Active sensing requires costly regulations compliance and good speech recognition

VCAs performing active sensing seem to be more frequent in academic research, while we could find only three such cases from industry, that is, LifePod, Pria, and WellBe. Although there are companies offering services to run market research surveys via voice assistants (e.g. True Reply, Voice Metrics), this practice is not yet extended to the field of mental and

substance use disorders. The main reason may be the significant costs of complying with HIPAA, making it harder for companies to invest in such a market (Chander et al., 2021).

Moreover, to interact with users, VCAs heavily rely on automatic speech recognition performance. To this day, different speech recognition methods have been explored (Bhatt, Jain, & Dev, 2021), yet factors like variation of user's voice, articulation, and background noise are still potential sources of error (Errattahi, El Hannani, & Ouahmane, 2018). Although individuals are capable of coping with speech recognition errors, these remain quite frequent (Myers, Furqan, Nebolsky, Caro, & Zhu, 2018). For example, imagine an individual who is having suicidal thoughts and wants to add an entry about it to a voice-based diary. The individual might tell the VCA with a shaky voice and fast-paced speech "I wanna die here." Given the poor quality of the utterance, the VCA might understand, for example, "I want a dryer." Consequently, the intent will be incorrectly stored, and the possible support intervention will fail to be correctly delivered. Moreover, voice interaction seems not to be as intuitive for all types of users. For instance, some older adults may face difficulty in producing an appropriate command or misunderstand how the VCA works (S. Kim, 2021). Hence, using VCAs for digital health interventions may require considerble speech recognition improvements.

10.4.4 Passive sensing requires extensive and rigorous data collection

According to the presented cases, vocal biomarkers monitoring via automated speech analysis is quite established. Why are VCAs not used for vocal biomarkers? If VCAs could sense when an individual is at risk of suffering from a given mental health condition or is showing signs of substance abuse, clinical outcomes may be detected promptly and allow patients to take measures early on. Possible reasons for VCAs not providing sensing solutions are difficulties in data collection, labeling, and feature analysis.

First, the VCA would need to be able to authenticate the patient correctly. It has been established that voice can be used to recognize the user but that noise may fool the recognition algorithm (Mohd Hanifa et al., 2021). Thus, for the algorithm to be accurate and precise, voice recordings would need to be the least polluted possible. This may constitute a challenge when using smart speakers to collect voice features from patients. Naturally, users are usually not right next to the device when interacting with it and ambient noise may compromise the quality of voice recordings. Furthermore, even assuming a collection of a clean audio signal, the recording still needs to be labeled to be used for the machine-learning model and to be later correctly analyzed for diagnostic purposes. Labeling during the training, validation, and optimization of the model requires human beings to annotate voice recordings to have accurate results. However, this is a demanding and time-consuming task. In addition, dealing with audio data also involves working with an enormous amount of data. Imagine that, for example, it requires over 630 MB of storage to for one hour of audio recording (with a ~44 kHz sample rate).

In addition, the application of vocal biomarkers may be problematic. That is, biomarkers may be biased by inter-individual variations depending on how the algorithms were trained. For instance, it has been observed that VCA users tend to produce shorter and possibly unnatural sentences, which are different from voice production during a human-to-human conversation (Pelikan & Broth, 2016). Moreover, the biomarker may be biased toward specific populations, whereas, for instance, the baseline values of features may be different between Argentine and Chinese women (Fagherazzi et al., 2021; Saggio & Costantini, 2020). Additionally, not only is detecting markers for specific health conditions heavily challenged by the high inter-individual variations, but intra-individual variations that are not related to changes in clinical outcomes make it difficult to detect changes in the health state (Anthes, 2020). For instance, dehydration (Alves, Krüger, Pillay, van Lierde, & van der Linde, 2019) or sleep deprivation (Icht et al., 2020) may also influence voice production. If a detection algorithm is not robust enough, it may influence clinical decisions contrary to the interest of the patient (Awaysheh et al., 2019). Thus, providing biomarkers based on vocal data may come with great responsibilities, depending on the gravity of the condition monitored. Hence, either the algorithm needs to be validated with the right ecological data and applied to corresponding populations and environments, or an important amount of data is required to make the vocal biomarker robust to demographic and contextual variations. These challenges may be observable in our cases: while academic research, which controls for sample demographics and recording environment, seems to confidently support the use of voice for mental health and substance use disorders detection, commercial solutions seem to be more cautious, whereas passive sensing is used to detect emotion in general, rather than specific health outcomes.

10.4.5 Concerns around data and conversation privacy

Recording voice naturally encounters ethical dilemmas around privacy protection as such as data may be used to infer identity and invade personal life (Fagherazzi et al., 2021). This is a problem, especially if the voice recordings are processed

and stored over the internet (e.g., cloud computing). Thus, encryption procedures are needed to ensure secure data storage and processing (Aloufi, Haddadi, & Boyle, 2019; Vaidya & Sherr, 2019).

Using VCAs for health may imply storing or retrieving personal and health data over voice and thus speaking aloud about potentially sensitive information. Although this may not be as problematic in private spaces (e.g., at home), users may be reticent to use VCAs for health in public spaces (Cowan et al., 2017; Moorthy & Vu, 2015). An extension to this problem is that VCAs may be hard to use for just-in-time adaptive interventions (Nahum-Shani et al., 2018), as a proactive voice-notification may be problematic if the VCA does not have context-awareness and users are not in a situation where they can receive it. An example may be while driving, a driver may be occupied with a maneuver or focusing on traffic, making the intervention more of a distraction than a help (Schmidt, Bhandare, Prabhune, Minker, & Werner, 2020; Schmidt, Minker, & Werner, 2020). One of the identified academic research cases (Adaimi et al., 2020) solved the receptivity verification using a wearable device. This is, however, technically less convenient as it requires coupling the VCA to an additional device.

Furthermore, while companies deliver solutions that are not illness-specific but that can be applied to both mental health and substance use disorders, academia tends to focus on selected conditions. In particular, academic research seems to focus either on evaluating existing commercial VCA for the support of specific health conditions (Boyd & Wilson, 2018; Miner et al., 2016; Nobles et al., 2020; Reis et al., 2018; Yang et al., 2021) or on developing Alexa skills for the same purpose (Cheng et al., 2018; Maharjan et al., 2019; Motalebi & Abdullah, 2018; Quiroz et al., 2020; Wang, Yang, Shao, Abdullah, & Sundar, 2020). In only a few of the reviewed cases, independent VCA prototypes were developed (Adaimi et al., 2020; Quiroz et al., 2020; Swamy et al., 2019). These were mainly intended as proof-of-concept solutions, which were not made available on the market. Thus, even though the technology may be more suitable for a specific condition, it remains unavailable to those in need. In comparison, the industry seems to diversify more in terms of implementation, hence making its technology available to different types of users while remaining less specialized. In particular, the interventions are either delivered as Google Actions and Amazon Skills to activate playback audio services (i.e., Calm, Headspace, Sleep Jar), which are easily accessible to Google and Amazon customers possessing a compatible device (e.g., smartphone or smart speaker), or they are delivered using services based on Amazon Alexa (i.e., Aiva, Orbita, Infermedica) or an independent product (i.e., Pria, WellBe, LifePod).

10.4.6 Amazon Alexa seems to rule the market

Nine of the presented cases describe a voice-based intervention delivered through commercial VCAs, using either Google Actions or Alexa Skills. Seven of those nine used Skills, whereas five were coming from academia (Cheng et al., 2018; Maharjan et al., 2019; Motalebi & Abdullah, 2018; Quiroz et al., 2020; Wang, Yang, Shao, Abdullah, & Sundar, 2020), and two were developed in industry (Headspace, 2021b; Jar, 2021). Only in a few cases, Google Actions were used (Calm, 2020; Headspace, 2021a; Infermedica, 2021; Jar, 2021). Moreover, two cases from industries dedicated to hospitals and aging facilities developed their service using the Amazon Alexa framework (Aiva, 2020; Orbita, 2020). Using existing voice technology services may simplify the implementation of health interventions but may also be problematic when it comes to health data storage, if the framework is not conforming to data protection regulations. In fact, in cases where personal and medical information is stored, either an independent VCA was developed (HandsFree, 2018; LifePod, 2019; Stanley Black & Decker, 2019), or Amazon Alexa was used (Aiva, 2020; Cheng et al., 2018; Maharjan et al., 2019; Motalebi & Abdullah, 2018; Orbita, 2020; Quiroz et al., 2020; Wang, Yang, Shao, Abdullah, & Sundar, 2020). Nevertheless, note that it has been observed that Amazon (together with Google) seems to be inconsistent in the process of Skill (or Action) approval (Cheng et al., 2020) suggesting that there is still a need for robust validation practices, especially in the health domain.

On the other hand, there are many alternatives to Amazon and Google, which are open-source and could allow developing services that are both HIPAA-compliant and more flexible. For instance, Kaldi (Kaldi ASR, 2021), VOSK (Alphacephei, 2021), and Julius (Julius, 2021) may be used for speech-to-text conversion, while ResponsiveVoice (Website) may perform text-to-speech conversion. Some of these (e.g., VOSK) allow for offline speech recognition, which may be beneficial for digital health interventions dealing with personal and health data.

10.5 Conclusion and outlook

This review aimed to present examples from academia and industry investigating or offering VCAs performing sensing and support for mental health and substance use disorders management. The primary goal was to provide cases of VCAs providing either passive sensing, active sensing, reactive support, proactive support, or a combination of these. As, to the best of the authors' knowledge, there is no VCA performing passive sensing, we included research and companies

focusing on vocal biomarkers for mental health and substance use disorders. As we observed that such vocal biomarkers are scientifically established, we assume that they could be used implemented in VCAs to provide just-in-time adaptive interventions. Moreover, we observe that most of the cases consist of reactive support solutions, while only a minority combines proactive and active support (i.e., Aiva), active sensing and reactive support (Quiroz et al., 2020), or active sensing with proactive support (i.e., LifePod, WellBe, Pria). Finally, we discuss possible reasons for VCAs not to fit in the sensing-and-support paradigm. In particular, we state that (1) reactive support may be the simplest way to assist individuals suffering from mental health or substance use disorders; (2) active sensing requires costly regulation compliance and may be hindered by voice recognition limitations; (3) passive sensing requires extensive and rigorous data collection; (4) concerns around health data sharing and privacy of spoken conversations about health matters may arise; (5) commercial framework solutions such as Amazon Alexa are most used when health data storage is required, as it simplifies and accelerates the implementation of interventions but not all Amazon Alexa solutions guarantee safe data management.

VCAs have enormous potential in relieving the healthcare system by providing automatized and standardized routine health interventions and helping individuals manage their mental health conditions or substance use disorders. This technology allows for accessible and efficient interaction, which is compatible with existing smart ecosystems (e.g., smart homes or cars). Moreover, VCAs represent the operationalization of the past and present enthusiasm in using human–computer conversations to control other devices more efficiently and monitor one's health. With the wide adoption of smart speakers and the ever-increasing use of VCAs for health-related purposes, digital health interventions may help individuals with mental and substance use disorders in a scalable way and at a low cost. Thus, we firmly believe future research should explore robust solutions allowing for a combination of sensing and support features in VCAs, to provide just-in-time health interventions.

References

104th United States Congress (1996). Health insurance portability and accountability act of 1996. *Public Law, 104*, 191.

Aarsland, D., Påhlhagen, S., Ballard, C. G., Ehrt, U., & Svenningsson, P. (2012). Depression in Parkinson disease—epidemiology, mechanisms and management. *Nature Reviews Neurology, 8*(1), 35–47. doi:10.1038/nrneurol.2011.189.

Abdolrahmani, A., Storer, K. M., Roy, A. R. M., Kuber, R., & Branham, S. M. (2020). Blind leading the sighted. *ACM Transactions on Accessible Computing, 12*(4), 1–35. doi:10.1145/3368426.

Adaimi, R., Ho, K. T., & Thomaz, E. (2020). Usability of a hands-free voice input interface for ecological momentary assessment. In *Paper presented at the 2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*.

Accenture (2020). Technology Vision 2020 | We, the Post-Digital People. https://www.accenture.com/_acnmedia/Thought-Leadership-Assets/PDF-2/Accenture-Technology-Vision-2020-Full-Report.pdf. Access date: 11 July 2022.

Alphacephei (2021). VOSK Offline Speech Recognition API. https://alphacephei.com/vosk/. Access date: 11 July 2022.

https://cogitocorp.com/evidence/. 2022. (Accessed 11 July 2022).

Agurto, C., Norel, R., Pietrowicz, M., Parvaz, M., Kinreich, S., Bachi, K., Cecchi, G., & Goldstein, R. Z. (2019). Speech markers for clinical assessment of cocaine users. *Proceedings of the ... IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP (Conference)*, *2019*, 6391–6394. https://doi.org/10.1109/icassp.2019.8682691.

Akçay, M. B., & Oğuz, K. (2020). Speech emotion recognition: emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication, 116*, 56–76. doi:10.1016/j.specom.2019.12.001.

Aiva. (2020). Aiva: Virtual Health Assistant. https://www.aivahealth.com/. Access date: 11 July 2022.

Alattas, A., Teepe, G., Leidenberger, K., Fleisch, E., Tudor Car, L., Salamanca-Sanabria, A., & Kowatsch, T. (2021). To What Scale Are Conversational Agents Used by Top-funded Companies Offering Digital Mental Health Services for Depression? In *Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC – 2021: Volume 5: HEALTHINF* (pp. 801–808). ISBN: 978-989-758-490-9 ISSN: 2184-4305.

Alves, M., Krüger, E., Pillay, B., van Lierde, K., & van der Linde, J. (2019). The effect of hydration on voice quality in adults: a systematic review. *Journal of Voice, 33*(1). 125.e113-125.e128 https://doi.org/10.1016/j.jvoice.2017.10.001.

Amante, D. J., Hogan, T. P., Pagoto, S. L., English, T. M., & Lapane, K. L. (2015). Access to care and use of the internet to search for health information: results from the US National Health Interview Survey. *Journal of Medical Internet Research [Electronic Resource], 17*(4), e106. doi:10.2196/jmir.4126.

Ammari, T., Kaye, J., Tsai, J. Y., & Bentley, F. (2019). Music, search, and IoT: How people (really) use voice assistants. *ACM Transactions on Computer-Human Interaction, 26*(3), 1–28. doi:10.1145/3311956.

Andreasen, N. C., & Grove, W. M. (1986). Thought, language, and communication in schizophrenia: diagnosis and prognosis. *Schizophrenia Bulletin, 12*(3), 348–359.

Anthes, E. (2020). Alexa, do I have COVID-19? *Nature, 586*(7827), 22–25. doi:10.1038/d41586-020-02732-4.

Aloufi, R., Haddadi, H., & Boyle, D. (2019). Emotionless: Privacy-preserving speech analysis for voice assistants. *arXiv preprint. arXiv:1908.03632*.

API - ResponsiveVoice.JS text to speech. (2015). https://responsivevoice.org/api/. Access date: 11 July 2022.

Apple (2021). Aiki - stress test & self care. https://apps.apple.com/us/app/aiki-stress-test-self-care/id1577209358. Access date: 11 July 2022.

Awaysheh, A., Wilcke, J., Elvinger, F., Rees, L., Fan, W., & Zimmerman, K. L. (2019). Review of medical decision support and machine-learning methods. *Veterinary Pathology, 56*(4), 512–525. doi:10.1177/0300985819829524.

Barata, M., Galih Salman, A., Faahakhododo, I., & Kanigoro, B. (2018). Android based voice assistant for blind people. *Libr. hi tech news, 35*(6), 9–11. doi:10.1108/lhtn-11-2017-0083.

Basatneh, R., Najafi, B., & Armstrong, D. G. (2018). Health sensors, smart home devices, and the internet of medical things: an opportunity for dramatic improvement in care for the lower extremity complications of diabetes. *Journal of Diabetes Science and Technology, 12*(3), 577–586. doi:10.1177/1932296818768618.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4*, 561–571.

Bedi, G., Cecchi, G. A., Slezak, D. F., Carrillo, F., Sigman, M., & De Wit, H. (2014). A window into the intoxicated mind? Speech as an index of psychoactive drug effects. *Neuropsychopharmacology, 39*(10), 2340–2348. doi:10.1038/npp.2014.80.

Bérubé, C., Schachner, T., Keller, R., Fleisch, E., V. Wangenheim, F., Barata, F., et al. (2021). Voice-based conversational agents for the prevention and management of chronic and mental health conditions: a systematic literature review. *Journal of Medical Internet Research [Electronic Resource]*. doi:10.2196/25933.

Bérubé, C., Kovacs, Z. F., & Fleisch, E. (2021). Kowatsch T Reliability of Commercial Voice Assistants' Responses to Health-Related Questions in Noncommunicable Disease Management: Factorial Experiment Assessing Response Rate and Source of Information. *J Med Internet Res, 23*(12), e32161. PMID: 34932003. doi:10.2196/32161.

Betthauser, L. M., Stearns-Yoder, K. A., McGarity, S., Smith, V., Place, S., & Brenner, L. A. (2020). Mobile app for mental health monitoring and clinical outreach in veterans: mixed methods feasibility and acceptability study. *Journal of Medical Internet Research, 22*(8), e15506. doi:10.2196/15506.

Bhatt, S., Jain, A., & Dev, A. (2021). *Continuous Speech Recognition Technologies—A Review* (pp. 85–94). Singapore: Springer.

Bostock, S., Crosswell, A. D., Prather, A. A., & Steptoe, A. (2019). Mindfulness on-the-go: effects of a mindfulness meditation app on work stress and well-being. *Journal of Occupational Health Psychology, 24*(1), 127–138. doi:10.1037/ocp0000118.

Boyd, M., & Wilson, N. (2018). Just ask Siri? A pilot study comparing smartphone digital assistants and laptop Google searches for smoking cessation advice. *Plos One, 13*(3), e0194811. doi:10.1371/journal.pone.0194811.

Bechtel, M., Briggs, B., & Buchholz, S. (2020). *Tech Trends 2020*. https://www2.deloitte.com/content/dam/Deloitte/ch/Documents/technology/deloitte-ch-Tech-Trends-2020.pdf. Access date: 11 July 2022.

Carolus, A., Binder, J. F., Muench, R., Schmidt, C., Schneider, F., & Buglass, S. L. (2019). Smartphones as digital companions: characterizing the relationship between users and their phones. *New Media & Society, 21*(4), 914–938.

Chander, A., Abraham, M., Chandy, S., Fang, Y., Park, D., & Yu, I. (2021). Achieving privacy: costs of compliance and enforcement of data protection regulation. *SSRN Electronic Journal*. doi:10.2139/ssrn.3827228.

Cheng, A., Raghavaraju, V., Kanugo, J., Handrianto, Y. P., & Shang, Y. (2018). Development and evaluation of a healthy coping voice interface application using the Google home for elderly patients with type 2 diabetes. In *Paper presented at the 2018 15th IEEE Annual Consumer Communications & Networking Conference (CCNC)*.

Cheng, L., Wilson, C., Liao, S., Young, J., Dong, D., & Hu, H. (2020). Dangerous skills got certified: measuring the trustworthiness of skill certification in voice personal assistant platforms. In *Paper presented at the Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event* https://doi.org/10.1145/3372297.3423339.

Cho, E., Molina, M. D., & Wang, J. (2019). The effects of modality, device, and task differences on perceived human likeness of voice-activated virtual assistants. *Cyberpsychology, Behavior, and Social Networking, 22*(8), 515–520. doi:10.1089/cyber.2018.0571.

Calm (2020). Calm for Google Home. https://blog.calm.com/blog/calm-for-google-home. Access date: 11 July 2022.

Choi, D., Kwak, D., Cho, M., & Lee, S. (2020). "Nobody Speaks that Fast!" an empirical study of speech rate in conversational agents for people with vision impairments. doi:10.1145/3313831.3376569.

Chung, A. E., Griffin, A. C., Selezneva, D., & Gotz, D. (2018). Health and fitness apps for hands-free voice-activated assistants: content analysis. *JMIR MHealth UHealth, 6*(9), e174. doi:10.2196/mhealth.9705.

Chung, S., & Woo, B. K. P. (2020). Using consumer perceptions of a voice-activated speaker device as an educational tool. *JMIR Medical Education, 6*(1), e17336. doi:10.2196/17336.

Cognitive Apps. (2021). Cognitive Apps. https://cogapps.com/. Access date: 11 July 2022.

Cohen, K. S. (1991). DragonDictate. *American Journal of Occupational Therapy, 45*(9), 856–857.

CompanionMx. (2018a). CompanionMx. https://companionmx.com/. Access date: 11 July 2022.

CompanionMx. (2018b). Evidence. https://companionmx.com/evidence/. Access date: 11 July 2022.

Corcoran, C. M., Carrillo, F., Fernández-Slezak, D., Bedi, G., Klim, C., Javitt, D. C., et al. (2018). Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry, 17*(1), 67–75. https://doi.org/10.1002/wps.20491.

Cowan, B.R., Pantidi, N., Coyle, D., Morrissey, K., Clarke, P., Al-Shehri, S., Earley, D. & Bandeira, N., 2017, September. "What can i help you with?" infrequent users' experiences of intelligent personal assistants. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services* (pp. 1-12).

Cummins, N., Epps, J., Sethu, V., Breakspear, M., & Goecke, R. (2013). Modeling spectral variability for the classification of depressed speech. In *Proc. Interspeech 2013* (pp. 857–861).

Damacharla, P., Dhakal, P., Stumbo, S., Javaid, A. Y., Ganapathy, S., Malek, D. A., et al. (2019). Effects of voice-based synthetic assistant on performance of emergency care provider in training. *International Journal of Artificial Intelligence in Education, 29*(1), 122–143. doi:10.1007/s40593-018-0166-3.

Dattani, S., Ritchie, H., & Roser, M. (2021). Mental Health. https://ourworldindata.org/mental-health. Access date: 11 July 2022.

David, E., & Selfridge, O. (1962). Eyes and ears for computers. *Proceedings of the IRE, 50*(5), 1093–1101.

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 28*(4), 357–366. doi:10.1109/TASSP.1980.1163420.

El-Shallaly, G. E. H., Mohammed, B., Muhtaseb, M. S., Hamouda, A. H., & Nassar, A. H. M. (2005). Voice recognition interfaces (VRI) optimize the utilization of theatre staff and time during laparoscopic cholecystectomy. *Minimally Invasive Therapy & Allied Technologies, 14*(6), 369–371. doi:10.1080/13645700500381685.

Errattahi, R., El Hannani, A., & Ouahmane, H. (2018). Automatic speech recognition errors detection and correction: a review. *Procedia Computer Science, 128*, 32–37.

Fagherazzi, G., Fischer, A., Ismael, M., & Despotovic, V. (2021). Voice for health: the use of vocal biomarkers from research to clinical practice. *Digital Biomarkers, 5*(1), 78–88.

Friedman, N., Cuadra, A., Patel, R., Azenkot, S., Stein, J., & Ju, W. (2019). Voice assistant strategies and opportunities for people with tetraplegia. doi:10.1145/3308561.3354605.

Gratch, J., Okhmatovskaia, A., Lamothe, F., Marsella, S., Morales, M., van der Werf, R. J., et al. (2006). Virtual rapport. In *Paper presented at the Intelligent Virtual Agents: 2006*.

Harandi, T. F., Taghinasab, M. M., & Nayeri, T. D. (2017). The correlation of social support with mental health: a meta-analysis. Electronic Physician,*9*(9), 5212–5222. doi:10.19082/5212.

HandsFree Health. (2018). Voice-enabled virtual health assistant. https://handsfreehealth.com/. Access date: 11 July 2022.

Headspace. (2021a). Headspace. https://www.headspace.com/. Access date: 11 July 2022.

Headspace. (2021b). Headspace on Alexa. https://www.headspace.com/alexa. Access date: 11 July 2022.

Headspace. (2021c). Headspace on Google Assistant. https://www.headspace.com/google-assistant. Access date: 11 July 2022.

Hoy, M. B. (2018). Alexa, siri, cortana, and more: an introduction to voice assistants. *Medical Reference Services Quarterly, 37*(1), 81–88.

Huang, L., Morency, L.-P., & Gratch, J. (2011). Virtual Rapport 2.0. In *Paper presented at the Intelligent Virtual Agents: 2011*.

Huang, Z., Epps, J., & Joachim, D. (2019). Investigation of speech landmark patterns for depression detection. *IEEE Transactions on Affective Computing, 1*. doi:10.1109/TAFFC.2019.2944380.

Huang, Z., Epps, J., Joachim, D., & Chen, M. (2018, September). *Depression Detection from Short Utterances via Diverse Smartphones in Natural Environmental Conditions* (pp. 3393–3397). INTERSPEECH.

IBM (2003). IBM archives: IBM shoebox. https://www.ibm.com/ibm/history/exhibits/specialprod1/specialprod1_7.html. Access date: 11 July 2022.

Icht, M., Zukerman, G., Hershkovich, S., Laor, T., Heled, Y., Fink, N., et al. (2020). The "Morning Voice": the effect of 24 hours of sleep deprivation on vocal parameters of young adults. *Journal of Voice, 34*(3). 489.e481-489.e489 https://doi.org/10.1016/j.jvoice.2018.11.010.

Infermedica. (2019a). Infermedica for Developers - Available conditions v3 https://developer.infermedica.com/docs/v3/available-conditions. Access date: 11 July 2022.

Infermedica. (2019b). Infermedica for Developers - FAQ. https://developer.infermedica.com/docs/faq. Access date: 11 July 2022.

Infermedica. (2021). Infermedica: Guide your patients to the right care. https://infermedica.com/. Access date: 11 July 2022.

Infermedica (2022). Symptomate. https://symptomate.com/diagnosis/. Access date: 11 July 2022.

Julius (2021). Open-source large vocabulary CSR engine Julius. https://julius.osdn.jp/en_index.php. Access date: 11 July 2022.

Kaldi, A. S. R. (2021). https://kaldi-asr.org/. Access date: 11 July 2022.

Kim, K., Boelling, L., Haesler, S., Bailenson, J. N., Bruder, G., & Welch, G. F. (2018). Does a digital assistant need a body? The influence of visual embodiment and social behavior on the perception of intelligent virtual agents in AR. In *Proceedings of the 17th IEEE International Symposium on Mixed and Augmented Reality (ISMAR 2018)* October 16–20, 2018.

Kim, S. (2021). Exploring how older adults use a smart speaker–based voice assistant in their first interactions: qualitative study. *JMIR mHealth and uHealth, 9*(1), e20427. doi:10.2196/20427.

Kinsella, B. (2020). *Smart speaker consumer adoption report 2020*. https://research.voicebot.ai/report-list/smart-speaker-consumer-adoption-report-2020/. Access date: 11 July 2022.

Kinsella, B., & Herndon, A. (2021a). *Smart speaker consumer adoption report 2021 – Germany*. https://research.voicebot.ai/report-list/germany-smart-speaker-consumer-adoption-report-2021/. Access date: 11 July 2022.

Kinsella, B., & Herndon, A. (2021b). *Smart speaker consumer adoption report 2021 – United Kindom*. https://research.voicebot.ai/report-list/uk-smart-speaker-consumer-adoption-report-2021/. Access date: 11 July 2022.

Kinsella, B., & Herndon, A. (2021c). *Smart speaker consumer adoption report – United States*. https://research.voicebot.ai/register/us-smart-speaker-consumer-adoption-report-2021/. Access date: 11 July 2022.

Kinsella, B., & Mutchler, A. (2019). *Voice assistant consumer adoptoin in healthcare*. https://voicebot.ai/wp-content/uploads/2019/10/voice_assistant_consumer_adoption_in_healthcare_report_voicebot.pdf. Access date: 11 July 2022.

Kubiak, T., & Smyth, J. M. (2019). *Connecting Domains—Ecological Momentary Assessment in a Mobile Sensing Framework* (pp. 201–207). Cham: Springer International Publishing.

Kulms, P., & Kopp, S. (2018). A social cognition perspective on human–computer trust: the effect of perceived warmth and competence on trust in decision-making with computers. *Frontiers in Digital Humanities, 5*, 14.

Large, D. R., Burnett, G., Anyasodo, B., & Skrypchuk, L. (2016). Assessing cognitive demand during natural language interactions with a digital driving assistant. In *Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 67–74). doi:10.1145/3003715.3005408.

Lattie, E. G., Adkins, E. C., Winquist, N., Stiles-Shields, C., Wafford, Q. E., & Graham, A. K. (2019). Digital mental health interventions for depression, anxiety, and enhancement of psychological well-being among college students: systematic review. *Journal of Medical Internet Research, 21*(7), e12869. doi:10.2196/12869.

Leikina, A. S. K. (2020a). *Cognitive Apps AI: research data*. https://cogapps.com/pdf/Cognitive%20Apps%20-%20Research%20Data%20(1).pdf.

Leikina, A. S. K. (2020b). *Cognitive apps emotion detection AI: accuracy testing overview*. https://cogapps.com/pdf/Cognitive%20Apps%20_%20Emotion%20Detection%20Accuracy%20Testing%20(2).pdf. Access date: 11 July 2022.

LifePod (2019). LifePod. https://lifepod.com/. Access date: 11 July 2022.

Low, D. M., Bentley, K. H., & Ghosh, S. S. (2020). Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investigative Otolaryngology, 5*(1), 96–116. doi:10.1002/lio2.354.

Maharjan, R., Bækgaard, P., & Bardram, J. E. (2019). "*Hear me out" smart speaker based conversational agent to monitor symptoms in mental health*. In *Paper presented at the Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*.

Masina, F., Orso, V., Pluchino, P., Dainese, G., Volpato, S., Nelini, C., et al. (2020). Investigating the accessibility of voice assistants with impaired users: mixed methods study. *Journal of Medical Internet Research [Electronic Resource], 22*(9), e18431. doi:10.2196/18431.

Meng, Z., Altaf, M. U. B., & Juang, B.-H. (2020). Active voice authentication. *Digital Signal Processing, 101*, 102672. doi:10.1016/j.dsp.2020.102672.

Militello, L., Sezgin, E., Huang, Y., & Lin, S. (2021). Delivering perinatal health information via a voice interactive App (SMILE): mixed methods feasibility study. *JMIR Formative Research, 5*(3), e18240. doi:10.2196/18240.

Miner, A. S., Milstein, A., Schueller, S., Hegde, R., Mangurian, C., & Linos, E. (2016). Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. *JAMA Internal Medicine, 176*(5), 619–625. doi:10.1001/jamainternmed.2016.0400.

Mohd Hanifa, R., Isa, K., & Mohamad, S. (2021). A review on speaker recognition: technology and challenges. *Computers & Electrical Engineering, 90*, 107005. doi:10.1016/j.compeleceng.2021.107005.

Moorthy, A., & Vu, K.-P. L. (2015). Privacy concerns for use of voice activated personal assistant in the public space. *International Journal of Human-Computer Interaction, 31*(4), 307–335. https://doi.org/10.1080/10447318.2014.986642.

Motalebi, N., & Abdullah, S. (2018). Conversational agents to provide couple therapy for patients with PTSD. In *Paper presented at the Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare*.

Mundt, J. C., Snyder, P. J., Cannizzaro, M. S., Chappie, K., & Geralts, D. S. (2007). Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. *Journal of Neurolinguistics, 20*(1), 50–64. doi:10.1016/j.jneuroling.2006.04.001.

Myers, C., Furqan, A., Nebolsky, J., Caro, K. & Zhu, J. (2018, April). Patterns for how users overcome obstacles in voice user interfaces. In *Proceedings of the 2018 CHI conference on human factors in computing systems* (pp. 1-7).

Nahum-Shani, I., Hekler, E. B., & Spruijt-Metz, D. (2015). Building health behavior models to guide the development of just-in-time adaptive interventions: a pragmatic framework. *Health Psychology, 34*(Suppl), 1209–1219. doi:10.1037/hea0000306.

Nahum-Shani, I., Smith, S. N., Spring, B. J., Collins, L. M., Witkiewitz, K., Tewari, A., et al. (2017). Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine, 52*(6), 446–462. doi:10.1007/s12160-016-9830-8.

Nahum-Shani, I., Smith, S. N., Spring, B. J., Collins, L. M., Witkiewitz, K., Tewari, A., et al. (2018). Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine: A Publication of the Society of Behavioral Medicine, 52*(6), 446–462. doi:10.1007/s12160-016-9830-8.

Namba, H. (2021). Physical activity evaluation using a voice recognition app: development and validation study. *JMIR Biomedical Engineering, 6*(1), e19088. doi:10.2196/19088.

Nass, C., & Lee, K. M. (2001). Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied, 7*(3), 171.

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. In *Paper Presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

Neurolex (2021). Neurolex. https://www.neurolex.ai/. Access date: 11 July 2022.

Nobles, A. L., Leas, E. C., Caputi, T. L., Zhu, S.-H., Strathdee, S. A., & Ayers, J. W. (2020). Responses to addiction help-seeking from Alexa, Siri, Google Assistant, Cortana, and Bixby intelligent virtual assistants. *npj Digital Medicine, 3*(1). doi:10.1038/s41746-019-0215-9.

Orbita (2020) Orbita AI – Leader in Conversational AI for Healthcare. https://orbita.ai/. Access date: 11 July 2022.

Oung, Q., Muthusamy, H., Lee, H., Basah, S., Yaacob, S., Sarillee, M., et al. (2015). Technologies for assessment of motor disorders in Parkinson's disease: a review. *Sensors, 15*(9), 21710–21745. doi:10.3390/s150921710.

Ownby, R. L., Crocco, E., Acevedo, A., John, V., & Loewenstein, D. (2006). Depression and risk for Alzheimer disease. *Archives of General Psychiatry, 63*(5), 530. doi:10.1001/archpsyc.63.5.530.

Perez Garcia, M., & Saffon Lopez, S. (2019). *Exploring the Uncanny Valley Theory in the Constructs of a Virtual Assistant Personality* (pp. 1017–1033). Cham: Springer International Publishing.

Pelikan, H. R., & Broth, M. (2016). Why that nao? how humans adapt to a conventional humanoid robot in taking turns-at-talk. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 4921–4932).

Pitardi, V., & Marriott, H. R. (2021). Alexa, she's not human but… Unveiling the drivers of consumers' trust in voice-based artificial intelligence. *Psychology & Marketing*. doi:10.1002/mar.21457.

Pradhan, A., Lazar, A., & Findlater, L. (2020). Use of intelligent voice assistants by older adults with low technology use. *ACM Transactions on Computer-Human Interaction*, (3373759). doi:10.1145/3373759.

Pradhan, A., Mehta, K., & Findlater, L. (2018). "Accessibility came by accident": use of voice-controlled intelligent personal assistants by people with disabilities. doi:10.1145/3173574.3174033.

Pure Tech (2022). https://puretechhealth.com/programs/details/sonde. (Accessed 11 July 2022).

Pulido, M. L. B., Hernández, J. B. A., Ballester, M. Á. F., González, C. M. T., Mekyska, J., & Smékal, Z. (2020). Alzheimer's disease and automatic speech analysis: a review. *Expert Systems with Applications, 150*, 113213. doi:10.1016/j.eswa.2020.113213.

Qiu, L., & Benbasat, I. (2005). An investigation into the effects of text-to-speech voice and 3D avatars on the perception of presence and flow of live help in electronic commerce. *ACM Transactions on Computer-Human Interaction, 12*(4), 329–355. doi:10.1145/1121112.1121113.

Quiroz, J. C., Bongolan, T., & Ijaz, K. (2020). Alexa depression and anxiety self-tests. In *Paper presented at the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and 2020 ACM International Symposium on Wearable Computers*.

Reis, A., Paulino, D., Paredes, H., Barroso, I., Monteiro, M. J., Rodrigues, V., et al. (2018). Using intelligent personal assistants to assist the elderlies An evaluation of Amazon Alexa, Google Assistant, Microsoft Cortana, and Apple Siri. In *Paper presented at the 2018 2nd International Conference on Technology and Innovation in Sports, Health and Wellbeing (TISHW)* https://doi.org/10.1109/TISHW.2018.8559503.

Saggio, G., & Costantini, G. (2020). Worldwide healthy adult voice baseline parameters: a comprehensive review. *Journal of Voice*. https://doi.org/10.1016/j.jvoice.2020.08.028.

Schmidt, M., Bhandare, O., Prabhune, A., Minker, W., & Werner, S. (2020). Classifying cognitive load for a proactive in-car voice assistant. In *Paper presented at the 2020 IEEE Sixth International Conference on Big Data Computing Service and Applications (BigDataService)*.

Schmidt, M., Minker, W., & Werner, S. (2020). How users react to proactive voice assistant behavior while driving. In *Paper presented at the Proceedings of The 12th Language Resources and Evaluation Conference*.

Reveri Health, 2022. Reveri: Digital Hypnosis. https://www.reveri.com/. Access date: 11 July 2022.

Schwartz, E. H. (2020). Coronavirus-related Google assistant actions blocked and removed. https://voicebot.ai/2020/03/09/coronavirus-related-google-assistant-actions-blocked-and-removed/. Access date: 11 July 2022.

Sezgin, E., Huang, Y., Ramtekkar, U., & Lin, S. (2020). Readiness for voice assistants to support healthcare delivery during a health crisis and pandemic. *npj Digital Medicine, 3*, 122. doi:10.1038/s41746-020-00332-0.

Sezgin, E., Militello, L. K., Huang, Y., & Lin, S. (2020). A scoping review of patient-facing, behavioral health interventions with voice assistant technology targeting self-management and healthy lifestyle behaviors. *Translational Behavioral Medicine, 10*(3), 606–628. doi:10.1093/tbm/ibz141.

Sezgin, E., Noritz, G., Elek, A., Conkol, K., Rust, S., Bailey, M., et al. (2020). Capturing at-home health and care information for children with medical complexity using voice interactive technologies: multi-stakeholder viewpoint. *Journal of Medical Internet Research, 22*(2), e14202. doi:10.2196/14202.

Shamekhi, A., Liao, Q. V., Wang, D., Bellamy, R. K., & Erickson, T. (2018, April). 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* (pp. 1–13).

Simmons, S. M., Caird, J. K., & Steel, P. (2017). A meta-analysis of in-vehicle and nomadic voice-recognition system interaction and driving performance. *Accident Analysis & Prevention, 106*, 31–43. doi:10.1016/j.aap.2017.05.013.

Sleep Jar (2021). Sleep Jar™ - sleep sounds & ambient noise. https://sleepjar.com/. Access date: 11 July 2022.

Sonde Health (2021). Sonde Health. https://www.sondehealth.com/. Access date: 11 July 2022.

Spitzer, R. L., Kroenke, K., & Williams, J. B. W. (1999). Patient Health QuestionnaireStudy Group. Validity and utility of a self-report version of PRIME-MD: the PHQ Primary Care Study. *JAMA, 282*, 1737–1744.

Spitzer, R. L., Kroenke, K., Williams, J. B., & Löwe, B (2006). A brief measure for assessing generalized anxiety disorder: the GAD-7. *Arch Intern Med, 166*, 1092–1097.

Spong, J., Graco, M., Brown, D. J., Schembri, R., & Berlowitz, D. J. (2015). Subjective sleep disturbances and quality of life in chronic tetraplegia. *Spinal Cord, 53*(8), 636–640. doi:10.1038/sc.2015.68.

Stacy M. Branham & Antony Rishin Mukkath Roy. 2019. Reading between the guidelines: How commercial voice assistant guidelines hinder accessibility for blind users. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '19)*. Association for Computing Machinery, New York, NY, USA, 446–458. https://doi.org/10.1145/3308561.3353797.

Stanley Black & Decker, Inc. (2019). Pria. Retrieved from okpria.com. Access date: 11 July 2022.

Strayer, D. L., Cooper, J. M., Goethe, R. M., McCarty, M. M., Getty, D. J., & Biondi, F. (2019). Assessing the visual and cognitive demands of in-vehicle information systems. *Cognitive Research: Principles and Implications, 4*(1). doi:10.1186/s41235-019-0166-3.

Stuart, S., & Koleva, H. (2014). Psychological treatments for perinatal depression. *Best Practice & Research Clinical Obstetrics & Gynaecology, 28*(1), 61–70. https://doi.org/10.1016/j.bpobgyn.2013.09.004.

Swamy, P. M., Janardhan Kurapothula, P., Murthy, S. V., Harini, S., Ravikumar, R., & Kashyap, K. (2019). Voice assistant and facial analysis based approach to screen test clinical depression. In *Paper presented at the 2019 1st International Conference on Advances in Information Technology (ICAIT)*.

Thakur, A., & Dhull, S. (2021). *Speech Emotion Recognition: A Review* (pp. 815–827). Singapore: Springer.

United Nations. (2021). *World Drug Report 2021*. https://www.unodc.org/unodc/en/data-and-analysis/wdr-2021_booklet-2.html. Access date: 11 July 2022.

Vaidya, T., & Sherr, M. (2019, May). You talk too much: Limiting privacy exposure via voice input. In *2019 IEEE Security and Privacy Workshops (SPW)* (pp. 84–91). IEEE.

Vtyurina, A., & Fourney, A. (2018). Exploring the role of conversational cues in guided task support with virtual assistants. doi:10.1145/3173574.3173782.

Wahsheh, L. A., & Steffy, I. A. (2020). *Using Voice and Facial Authentication Algorithms as a Cyber Security Tool in Voice Assistant Devices* (pp. 59–64). Cham: Springer International Publishing.

Wang, L., & Miller, L. C. (2020). Just-in-the-moment adaptive interventions (JITAI): a meta-analytical review. *Health Communication, 35*(12), 1531–1544. doi:10.1080/10410236.2019.1652388.

Wang, J., Yang, H., Shao, R., Abdullah, S., & Sundar, S. S. (2020, April). Alexa as coach: Leveraging smart speakers to build social agents that reduce public speaking anxiety. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1-13).

Yang, S., Lee, J., Sezgin, E., Bridge, J., & Lin, S. (2021). Clinical advice by voice assistants on postpartum depression: cross-sectional investigation using apple Siri, Amazon Alexa, Google Assistant, and Microsoft Cortana. *JMIR mHealth and uHealth, 9*(1), e24045. doi:10.2196/24045.

Yeager, C. M., & Benight, C. C. (2018). If we build it, will they come? Issues of engagement with digital health interventions for trauma recovery. *mHealth, 4*, 37. doi:10.21037/mhealth.2018.08.04.

Young, R. A., & Zhang, J. (2017). Driven to distraction? a review of speech technologies in the automobile. *The Journal of AVID*. http://acixd.org/wp-content/uploads/2018/10/2-Young-2017-Speech-rev.-29-6-9-17.pdf. Access date: 11 July 2022.

Non-Print Items

Abstract

While mental and substance use disorders are worryingly prevalent worldwide, voice-based conversational agents (VCAs) are penetrating our homes and are increasingly used for health-related purposes. As voice interaction is less effortful and more accessible than visual interfaces, VCAs may constitute a scalable solution for the delivery of health interventions. In particular, VCAs may respond to a requenst for support, as well as proactively providing it to the user. Also, voice interaction may be used to timely sense critical health states by gathering questionnaire data, as well as passively collecting acoustic data streams associated with health-related variables. In this chapter, we review what is currently available for individuals with mental and substance use disorders through the lens of a *sensing-and-support paradigm*. In particular, we present examples of VCAs and voice technology from academia and industry, and identify current capabilities and potentials for the management of mental and substance use disorders. Furthermore, we seek to illustrate the implementation gaps in comparison to the sensing-and-support paradigm and discuss the possible reasons for such gaps (i.e., implementation, cost, data management, and privacy concerns).

Keywords

Digital health; mHealth; Voice assistant; Smart speaker; Smartphone; Academia; Industry