Conversational User Interfaces on Mobile Devices: Survey

Razan Jaber
DSV, Stockholm University
Stockholm, Sweden
razan@dsv.su.se

Donald McMillan
DSV, Stockholm University
Stockholm, Sweden
donald.mcmillan@dsv.su.se

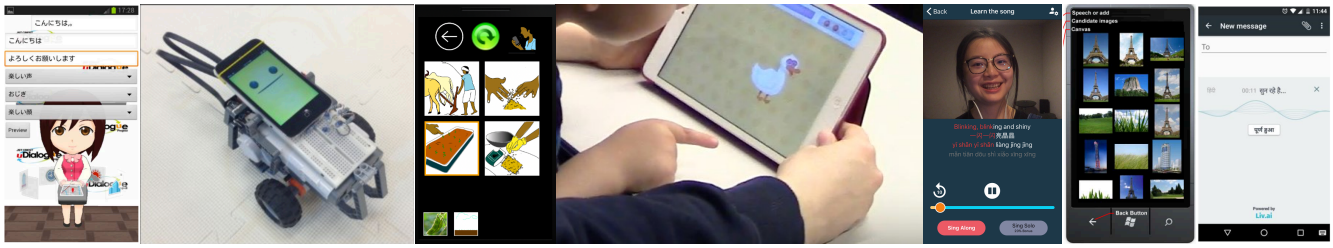

Figure 1: A Selection of Speech-based Interfaces on Mobile Devices.
Mei-chan [96], Quinn [61], VideoKheti [22], Cookie Monster's Challenge [18], SLIONS [68], JIGSAW [93], & Swarachakra [9]

ABSTRACT

Conversational User Interfaces (CUI) on mobile devices are the most accessible and widespread examples of voice-based interaction in the wild. This paper presents a survey of mobile conversation user interface research since the commercial deployment of Apple's Siri, the first readily available consumer CUI. We present and discuss *Text Entry & Typing*, *Application Control, Speech Analysis, Conversational Agents, Spoken Output*, & *Probes* as the prevalent themes of research in this area. We also discuss this body of work in relation to the domains of *Health & Well-being*, *Education*, *Games*, and *Transportation*. We conclude this paper with a discussion on *Multi-modal* CUIs, *Conversational Repair*, and the implications for CUIs of greater access to the *context* of use.

CCS CONCEPTS

• **Computing methodologies** → *Speech recognition*; • **Human-centered computing** → **Interaction techniques**; *Smartphones*; • **General and reference** → **Surveys and overviews**.

KEYWORDS

Conversational User Interfaces, Smartphones, Voice Interaction, Agent Interaction, Literature Survey

ACM Reference Format:
Razan Jaber and Donald McMillan. 2020. Conversational User Interfaces on Mobile Devices: Survey. In *2nd Conference on Conversational User Interfaces (CUI '20), July 22–24, 2020, Bilbao, Spain.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3405755.3406130

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CUI '20, July 22–24, 2020, Bilbao, Spain
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7544-3/20/07…$15.00
https://doi.org/10.1145/3405755.3406130

1 INTRODUCTION

The most common deployment of Conversational User Interfaces (CUIs) are those on mobile devices. From Apple's Siri to Google's Assistant, most mobile operating systems come with access to at least one conversational user interface in the form of an interactive personal agent. Despite this, the majority of research on CUIs seems to focus elsewhere, with only 15% of papers referenced in the recent review by Clark et al. [19] focusing on mobile interaction even after excluding embodied interaction with devices such as stand-alone smart speakers and robots. In this paper we present a survey of research on mobile conversational user interfaces, examining the opportunities and challenges in developing them for this form-factor.

The motivation is to explore how mobile devices offer specific challenges and opportunities for CUIs, which can inform the design of CUIs more broadly. One of the largest of the opportunities is scale. Taking the U.S. as an example, in 2019 it was estimated that 26% of consumers had access to a smart speaker [51] compared with 86% having access to a smart phone [86]. However, taking advantage of this opportunity requires understanding the difference between developing a CUI on a standalone smart speaker or an embodied robot and developing for a consumer mobile device. We examine how the touch screen is, and can be, taken advantage of alongside CUIs, how the personal nature of a modern mobile device can be leveraged, and how the multitude of sensors on these devices can be used to inform CUIs of issues such as the physical and social contexts of the user to provide a more situated and responsive interaction.

For this reason we focus the scope of the survey presented here to CUIs on mobile devices. By searching using a combination of the keywords from the proceedings of the 2019 1st annual Conference on Conversational User Interfaces (CUI'19), we present a structured overview of recent research in this area. We discuss this body of work in relation to the dominant themes which are distinctive to the form factor on which the CUIs were developed or studied.

	References
Themes	
Text entry	[9, 38, 67, 71, 82, 85, 93]
Application control	[22, 90, 92]
Speech analysis	[26, 36, 52, 68, 72]
Spoken output	[28, 31, 66, 77]
Conversational Agents	[33, 43, 45, 48, 55, 61, 96]
Probes	[18, 74]
Domains	
Education	[22, 43, 52, 61, 68, 77]
Health & Well-being	[26, 28, 36, 71, 72, 92, 94]
Transportation	[33, 38, 55, 67, 90]
Gaming	[18, 26, 36, 52, 66, 68, 72]

Table 1: Themes and Domains

Following this, we discuss four prevalent domains of application of CUIs on mobile devices. We end with a higher level discussion of the impacts of multi-modality on CUIs, how the form-factor can influence the recovery from breakdowns in communication between human and device, and how the mobile device can aid in the contextualisation of CUI interaction.

2 METHODOLOGY

In selecting papers for this survey, a comprehensive literature review was carried out based on the collected keywords of all published papers in the proceedings of CUI'19, the 1st International Conference on Conversational User Interfaces. After accounting for duplicates, keywords that were too general (e.g., trust, privacy, design), and keywords in different fields (e.g., performing arts, AI governance, emergency calls) we were left with 36 optional keywords [1], to which we added two that were required: "speech" and "interaction." These were added to ensure that the papers included speech interaction and not only research on the detection or the production of speech. We chose a date range of the period after 2011, when Apple announced Siri on the iPhone – the first widely available CUI on consumer mobile devices.

We used these criteria to search the ACM digital library using the full 'ACM Guide to Computing Literature' database which indexes over 6,500 other publishers' content including IEEE Computer Society and Springer-Verlag[2]. The search was performed through a full text search in any field, capturing publications that mention these terms anywhere in the paper. This search yielded 3014 research papers. These papers were then further filtered by two independent coders reading each abstract (with a 12% overlap) and applying the following relevance criteria:

(1) The paper describes a speech interface that is deployed on a mobile device.
(2) The paper describes a system that uses speech as input, output, or both.
(3) The paper covers an evaluation process that focuses on the use of the technology.

We have excluded technical discussions of systems, models, or methods related to speech interfaces that had little to no user evaluation. We also excluded extended abstract publications, workshop papers, magazine articles, and panel discussions. After the abstract screening, the remaining 45 papers were read in full, resulting in a further 17 being excluded. The resulting 28 papers are included in the final analysis.

The analysis centred around the authors reading and independently categorising the papers in terms of theme, domain, interaction modalities, contribution, and study design. The authors engaged in repeated rounds of analysis, which followed thematic analysis practice of developing codes independently, combining them collaboratively, and identifying candidate themes, domains, and their connections. The papers were finally arranged for presentation based on the *Interaction Themes* and *Application Domains* of the system under investigation. In developing the themes, the focus was on connecting the interaction with the CUI to the salient characteristics of mobile interaction [14, 17].

For the six themes (**Table 1, Top**), we present the included papers by describing them in relation to the interface (e.g., the type of other modalities used along with speech) most relevant to the contribution of that work. The themes cover all papers selected for analysis in one, most relevant, theme. Following this, the four application domains (**Table 1, Bottom**) are presented in relation to the opportunities and challenges of speech interaction in those domains. In the domain categories, some studies are aligned to more than one domain, and for others, the domain was either not relevant or not frequent enough for detailed discussion.

3 THEMES

In this section, we describe the surveyed work in relation to the themes identified in **Table1**. This allows us to focus on the opportunities presented by these types of interactions. Each section provides a brief overview of the theme in relation to the research presented and concludes with a discussion of its relevance to CUIs and mobile devices.

3.1 Text Entry and Typing

Given the amount of time the user spends entering text using their smartphone keyboard [14, 15], employing CUIs for effective text-entry can have a significant impact on the quality and efficiency of the user experience of the device. Several studies have been conducted to compare typing and text entry using speech-based interfaces in both 'hands-free' and 'eye-free' configurations. Munger et al. [67] conducted a study to compare destination entry for a navigation application using a keyboard and speech-based interaction with and without a wake word. The study has shown that voice interfaces provide a significant advantage over the touch interface in terms of distraction and length of interaction while driving. Similarly, He et al. [38] also conducted a study to asses the

[1] chatbot, conversational interface, speech interface, voice user interface, intelligent personal assistant, conversational agent, conversational user interface, voice assistant, anthropomorphism, conversation analysis, conversational AI, social robot, dialog systems, smart speaker, speech synthesis, voice, agents, Emotion, ethical decision making, conversation, personification, robot, NLP, speech technology, voice interaction, dialog design, ethnomethods, conversational design, intentional interface, multimodal interaction, face-to-face conversation, embodied communication, gaze interaction, text-to-speech, embodiment

[2] https://libraries.acm.org/digital-library/acm-guide-to-computing-literature

effects of typing or speech input on driving performance by asking participants to send a text message, either using a CUI or through a keyboard. They found that handheld texting increased the brake response time, among other safety-related factors, in comparison to the CUI use. This gives evidence that speech-based technologies can reduce manual and visual interference when compared to using touch screen in driving context. A comparison study conducted by Bhikne et al. [9] found that the combination of speech input with a keyboard for error correction was more than 2.5 times faster than using only the keyboard. Ruan et al. [82] argue that speech-based text entry in English could be almost 3x times faster than keyboards.

In a different area, users have used speech-based interaction to enhance usability. Jacko et al. [71] designed a multilingual communication support system to improve the communication between hospital staff and foreign patients. The system outputs a translated sentence when users input sentences to be translated with a combination of spoken, touch screen selection, and keyboard input.

An interactive mobile image search application has been presented in [93] to facilitate visual search on mobile devices, JIGSAW. The application uses voice input to initiate the query, which can then be manipulated at a more granular level using other input modalities, which the authors argue is an effective, complementary, interactive paradigm for mobile search. Mobile CAPTCHA entry has also been investigated, with Shirali-Shahreza et al. [85] allowing users to use speech to complete a CAPTCHA instead of typing it. Their results showed that participants preferred to speak their interpretation of the CAPTCHA rather than typing it, even with the associated possible transcription errors, although they preferred to read rather than having it spoken for the same reason.

The studies here have shown that speech input works better in scenarios when there is a small amount of text to be entered, such as SMS dictation. Considering that speech recognition is not 100% accurate, some techniques to correct errors are needed, and the form-factor of the modern smartphone provides opportunities for novel supporting methods – from simply taking advantage of the screen to display the results of the transcription, to developing novel iterative and interactive methods to confirm and refine speech-based text entry without directly having to resort to a keyboard. We do not see evidence that CUIs can, or should, replace keyboards for all tasks on mobile devices. However, the opportunities for complementing and collaborating with keyboard-based input present themselves as an interesting avenue for future research at the confluence of CUI, HCI, and UX.

3.2 Application Control

Speech interfaces have also been used successfully to control and launch applications on smartphones without necessarily relying on presenting the transcription of a query to the user for confirmation.

Cuendet et al. [22] used speech in combination with the smartphone touchscreen to design the interface of VideoKheti, an instructional video repository for farmers in rural India. The interface is completely text-free, with a "push-to-talk" button to allow users to find the informational video that met their needs through selecting from images or stating the name of their choice on the current level of the navigation tree. Here the authors used the SALAAM method [75, 77] to train a small vocabulary recogniser in the local dialect with minimal training data [22]. By comparing the use of the system with and without speech-enabled, they saw clear benefits, especially to those with lower literacy levels, even though the interface itself was text free in both conditions. One problem with having the interface text-free was that to provide enough detail and cover enough situations for the application to be useful resulted in a relatively large specialised vocabulary, and long spoken lists, which were seen to overload users. The application did, however, provide a series of graphical 'breadcrumbs' as part of the interface allowing the user to keep track of where they were in the decision tree and backtrack as necessary.

Tchankue et al. [90] performed a usability study on an in-car speech interface, but without the opportunity to dictate text to be sent as a message. The interface they tested allowed users to either call or send pre-determined text messages (from a list of 4 to make it easy to remember) to someone on their contact list, or to a dictated phone number. With the inclusion of 'cancel' and 'repeat' commands, they found that even navigating this simple application through voice was susceptible to recognition errors, and that the users sorely missed a barge-in feature wrestle back control from the system.

In the accessibility context, Wang et al. [92] proposed EarTouch, a one-handed eyes-free interaction technique that allows the users to interact with a smartphone using their ear to perform tap or draw gestures on the touchscreen. Alongside this, they were also able to use voice commands, with listening triggered using their ear, to specify a map destination, or and to send voice messages.

The opportunities for CUI exposed by the work under this theme take advantage of the hands and eyes-free nature of much of spoken interaction with mobile devices. This, echoed in the driving domain, shows that the trade-offs involved in moving from touch to spoken interaction can be easily justified for reasons of safety. Another application type in this theme focused on accessibility, specifically providing the opportunity for low-literate users to better interact with the traditional text-based mobile applications. The combination with a GUI presents opportunities to ameliorate some of the most fundamental problems with CUIs; awareness of the current state of the system and the shared contextual model of interaction (by using 'breadcrumbs' for example), and problems resulting from the awareness of the extent and contents of large lists of options available to the user.

3.3 Speech Analysis

In this theme many of the studies present speech interfaces that aim to analyse users' utterance either for language learning or therapy, rather than to directly use *what* they say.

In the context of learning languages, Kumar et al. [52] explored the use of speech recognition to help children in rural areas read and understand words in their local dialect, with recognition isolated to individual words. They designed an educational game where the player was shown an image and had to say the corresponding word. If the Automatic Speech Recognition (ASR) system was able to understand them, they were regarded as speaking the correct word. The game was designed to encourage repeated vocalising to cement learning. Several studies went further than checking if the

word was understood or not and included the ability to distinguish between different ways in which incorrect pronunciations were voiced to give specific feedback.

Parnandi et al. [72] also created an application that required the users to produce utterance in response to images displayed on the screen, this time with a focus on speech therapy for children with apraxia of speech. In this case, the system was able to identify insertion, deletion, and substitution mispronunciations on a phoneme by phoneme level. On displaying this to the remote therapist, they were able to assign or design speech exercises for the children based on their specific difficulties. SpokeIt [26] by Duval et al. is another speech therapy game, capable of providing real-time feedback to help children with speech impairments. The goal was to detect miss-pronunciation, and to that end, they used a custom dictionary in Pocketsphinx [41] encoded with common miss-pronunciations of the words in the training exercises which were then able to be analysed and presented to therapists or parents. The system allowed therapist to remotely follow their progress and assign specific speech production exercises to each child depending on what was needed for that child. It also helped the learners by demonstrating the correct pronunciation with lip animations. Hair et al. [36] presented a third remote therapy tool. They used a Wizard-of-Oz trial to compare the potential for improving engagement with therapy through the use of a mobile game to motivate practice.

Outside of the therapy domain Murad et al. [68] crated a multilanguage karaoke application called SLIONS to aid in learning to pronounce a new language. This application matched pronunciation between the user's singing and the 'correct' pronunciation of the song. This allowed for personalised granular feedback on their singing pronunciation. The study showed that the system helped to improve users' vocabulary and pronunciation.

One important ability that machine-based speech interfaces are able to provide a detailed and granular representation of the utterances that the users produce automatically, which are then available to be analysed or stored and compared over time to detect improvement or degradation of ability. In this theme, the applications all provide the users with some form of feedback on *how* the utterance was spoken, for use in both improving pronunciation for language learning and providing tools for a variety of speech therapy regimens. In speech therapy and language learning applications, the use of the screen in addition to the CUI allowed the exercises to vary beyond 'call and repeat' drills. The personal nature of most mobile devices, even for children, can also be important in comparison to a 'family' smart speaker. This provides reassurance that the dissemination of the personal feedback information can be managed by the end user in the same way they manage access to their device or other overhearings, and that the therapist can ensure that the correct user receives the training regime that best suits them individually.

3.4 Conversational Agents

Conversational agents have been employed on mobile devices to both provide access in specific locations and to take advantage of the touchscreens to enhance the interaction. Lubold et al. [61] introduced a socially responsive interface called 'Quinn' which had a simple face-avatar displayed on the mobile screen (see **Figure 1: second from the left**) and which adapted the prosody of its responses based on the pitch of the user's utterance. They argue that adaptive, socially responsive speech interfaces can be beneficial beyond the education domain they focused on. Yamamoto et al. [96] presented Mei-chan (see **Figure 1: first from the left**), an animated always-on virtual agent on a mobile device with the goal of a more natural interaction by using a combination of speech manipulation and animation to communicate emotion. Similarly, Kang et al. [48] produced an animated avatar for their conversational agent with upper body movements and facial expressions. This allowed them to experiment with the impact that mutual gaze and gaze aversion had on agent interaction.

Gordon et al. [33] introduced a parental driving entertainment virtual agent called PANDA connecting two tablets, one in the front, and one in the back of the car. The agent supports the parent driver via a CUI to mediate interaction with children in the back seat. Large et al. [55] conducted an investigation study to asses the cognitive demand or carrying out natural language interactions with a digital driving assistant while driving. They found a level of secondary task completion equivalent to conducting a conversation with another person using a hands-free mobile phone.

Jain et al. [44] designed a conversational agent that provides farming information via a CUI, called FarmChat. They conducted an evaluation focusing on assessing the usability of the system, and understanding people's needs and challenges in using voice-based interfaces in rural locations. Jarusiboonchai et al. [45] also took advantage of the mobility of the mobile phone to embed their Wizard-of-Oz based trial of a proactive conversational facilitator between co-located participants in different settings.

The smartphone provides opportunities for both the application and design of conversational agents. Animated agents can be an effective way to provide nonverbal expression (such as facial expressions, head nods, eye gaze, and posture shifts), and increase communicative engagement with users. Taking advantage of these animations can facilitate emotionally engaging social interactions [49]. The screen provides a high fidelity output for complex animations, which can be incorporated in agent output. The mobile also provides direct actions that the agent can take, increasing their perceived usefulness and agency, such as playing multimedia content or sending messages on behalf of the user. The ability to be *in situ* with the user in a specific context should also be explored, as location or interlocutor awareness can also enable actions and reactions beneficial to the ongoing engagement with the agent. By using other sensors on the phone, for example, the camera, the CUI could also understand subsets of facial expressions, gaze, and gestures for more natural and intuitive communication.

3.5 Spoken Output

A number of systems focused only on generating spoken output for specific purposes. El-Glaly [28] developed an intelligent reading support system for blind people. It harnessed spatial, auditory, and haptic feedback to follow the user's finger across text, which it would read to them. Raza et al. [77] introduces two further systems with simple voice navigation menus. Both of their systems, 'Songline' for sharing music and 'Polly' for manipulating and sharing

spoken messages, allowed the users to record, share, find, and listen to audio through menu trees of spoken option, with input from the keypad on the phone. The focus was to use viral media sharing as entertainment as a way to train users on speech-enabled technologies implicitly. Moran et al. [66] integrated text-to-speech with a pervasive mobile game called 'Cargo' to provide instructions to multiple teams of players. Fiannaca et al. [31] looked to provide expressivity to the speech that generated from text-to-speech engines through a touch screen interface. The model allows expressivity through the insertion of emoji, punctuation, vocal sound effects like laughter into the synthesised speech, which was highly valued by end-users.

The generated speech could also be augmented with voice adaptation based on the state of the interaction, for example the speed of touch interaction could be mirrored in the pitch or the speed of the generated speech. With the addition of a touch screen, camera, or a simple keypad, speech output can be disconnected from ASR and embedded in other aspects of mobile device interaction.

3.6 Probes

This theme comprises of two papers, where a technology probe using and about CUIs was deployed. Cheng et al. [18] examined conversational repair strategies to correct communication breakdowns with a voice-driven interface. In this study, they have examined the repair strategies [7] that young children face when they play a game using voice interaction, and the system deliberately fails to recognise their utterance. They have found that young children borrow conversational strategies that are common in person-person interactions and applied them to their attempts to interact with the app such as, raising their voice. Porcheron et al. [74] used video analysis to examine the use of CUI-based personal assistants on mobile devices (such as Apple's Siri, or the Google Assistant) in the social setting of groups meeting in a cafe. Among other things, they looked at how the break in the ongoing conversation necessary for the CUI to recognise the query is negotiated between the conversational partners and how repetition and breakdowns are handled in conversation.

These papers highlight that using a conversational agent deployed as technology or cultural probe is a highly underutilised application of CUI. While there have been some work on probes *for* potential CUI work, such as [62], and extensive work on cultural probes that employ digital photography or audio recording [11, 34, 42] the work done by Cheng et al [18] and Porcheron et al. [74] can be seen as an interesting turn towards using mobile CUIs for examining both people interacting with speech-based systems, but also how people act around, and outwith the technology we tend to focus on. Adapting mobile CUIs to be proactive in their interactions with participants in combination with the awareness of the mobile devices' contexts holds a further opportunity to probe specific social and cultural practices using natural conversation, instead of relying on participants to editorialise their experiences in textual or visual form, or submit to more invasive observational methods.

4 DOMAINS

In addition to the themes, between which all papers are distributed, we have also identified a number of cross-cutting domains, which can also be discussed in terms of the application and opportunities for CUIs. In this section, the domains shown in **Table 1** of Health & Well-being, Education, Transportation, and Games are discussed.

4.1 Health & Well-being

One interesting possible application of CUIs in healthcare is collecting patient medical information without, or before, interacting with a doctor. Many phone or video-based Telehealth systems provide the opportunity to collect medical history, report symptoms, and receive a non-urgent diagnosis or arrange for followup consultations [25, 91]. While this provides wider and more convenient access for some, it still requires highly trained healthcare professionals available in real-time. There have been studies highlighting the potential benefits of using conversational agents for health-related purposes, for example, assisting clinicians during the consultation or assisting patients in compiling health-related diaries [54]. Combining these trends in mobile CUIs presents as an impactful and fruitful opportunity for research. CUIs can be employed to support patients in self-management by providing reminders, answering common questions on ongoing conditions or diagnoses (dates of planned treatment, doses of medication, etc.), or supporting in self-monitoring by vocally prompting for health diary entries. These systems can also lead the user through more complex conversational trees to adapt its contents to fit the changing health situation of the patient over time. Using mobile CUIs in this domain could decrease health care costs, and make it easier and more convenient to deliver therapy to those in need. There is also the possibility to combine such system with the already complex data collection systems embedded in mobile devices, allowing the application to collect and share health data when needed. The sensors on the mobile device allow a mobile application to collect information about the user's activities, preferences [95], and the surrounding environment [60]. Sound captured by a mobile phone's microphone is a rich source of information that can be used to make more accurate inferences about the person carrying the phone, their environment, and social context [64]. Indeed, smartphones and related digital sensing devices already offer data collection, activity suggestions [30], and reporting for safety purposes in the eldercare domain [70]. Enhancing these systems with CUIs should be an expected development in the near future.

Using virtual agents that can approximate face-to-face interaction with the patient could help in health care communication. By including simulations of nonverbal conversational behaviour, including hand gestures, facial displays, posture shifts, proxemics, and gaze, this additional modality could be important for establishing trust, rapport, and therapeutic alliance [40] with patients. Health care education systems use such multi-modal CUIs to provide personal mobile virtual agents to act as coaches able to educate and counsel patients on a variety of topics. They also offer the opportunity to mitigate some of the isolation of patients with mental health issues, such as depression, [65]. Such personal communication CUIs can also provide an opportunity to detect when a user needs urgent interventions [69] while reaching populations who

report fear of the stigma associated with seeking professional services for sensitive mental health issues [79]. However, providing CUIs that provide an approximation of emotional output raises the expectation that the agent would also have some understanding of the user's emotional state using the same cues. Even as the field moves beyond basic sentiment analysis to more sophisticated techniques, possibly taking advantage of the mobile phone's access to the user's context, such emotional understanding is still in its infancy. This balance between engendering trust and over-promising the emotional and cognitive abilities of the agent there should be given even greater considerations designing when CUIs for the health care domain, almost to the same level as issues of accuracy, privacy, and confidentiality [10] are given currently.

4.2 Education

Education using mobile phones hold the promise of facilitating learning, promoting collaboration, and encouraging both independent and cooperative learning for life. In combination with CUIs, mobile learning can be offered with direct access to information without the need to navigate through a GUI or complex menu structure, which is well-suited for low-literate users, as it leverages a skill they already have – this can even, to some extent, mitigate the challenges of living with low-literacy by providing access to information and services they would otherwise be excluded from. Beyond those with low-literacy, students all over the world experience personal learning through mobile devices. Didactic applications and interfaces on smartphones are used both at a distance and during face-to-face studies. They may, moreover, be used to bring new ways of developing generic and specific competencies for different users. Using CUIs integrated with current didactic systems, or specially designed teaching CUIs on mobile devices offer personal multi-modal opportunities for learning.

When looking at the use of CUI in the educational domain, there is a research focus on how these systems have been used by children [52] and adults [87] to learn with these technologies beyond institutional contexts (i.e., informal learning) and their implications on learning.

The most important feature of mobile phone technologies is their portable nature and their abilities to promote additional learning methods beyond the classroom. Smartphones provide learning and training support for students enabling quick content delivery, enhanced support time in project-based group work, a higher level of student engagement in learning-related activities within a multitude of diverse physical locations, and the enhanced availability and accessibility of information. Combining this with the advantages of CUI based learning, for example, of increased motivation and promoting learning-through-teaching, can provide customised solutions that foster successful learning and performance without instructor, where the user can interact with the system using their natural language. Designing educational CUIs for mobile devices allows developing useful customisation that exploits adaptive instant interaction based on context with real-time responses. Conversational systems can initiate natural language queries and expect to receive the learner's natural language response [84]. The system would form a dialog-oriented intersection between a human and the system that allows for natural communication, with real-time responses feedback. Automatic speech recognition (ASR) has been used in many applications specifically for language training, vocabulary practice, and for improving pronunciation [21, 29, 39].

Using CUIs in educational applications could compensate the lack of face-to-face interactions, which have been shown to give the students more motivation [46], by increasing the interaction with the learner enabled through a human-like tutor agent. Conversational agents with simulated human-like interfaces can be used to facilitate interactions between the learner and the content, where the role of the agent in this case is to present the instructional content [87]. Through the use of a conversational agent the learner can participate in reflection activities [20] which can be guided by the instructor through their crafting of the dialogue – but would not require their immediate presence. Taking advantage of the mobile device here allows for this reflection to be based upon both personal data (such as individual learning goals or notes), collaborative data from the learner's cohort as a whole, and resources accessed through the network.

4.3 Transportation

When looking at the use of conversational user interfaces in the transportation domain, the most commonly explored goal is that of reducing the level and amount of distraction experienced by drivers. The task of driving places demands on visual attention – both towards the road and other road users outside of the vehicle, and the variety of visual feedback provided inside the vehicle on its ongoing state. At the same time it is a manual task, involving frequent control of the direction and speed of the vehicle through physical manipulation of the steering wheel and pedals. It is also a cognitive task [23], as drivers plan and re-plan their manipulations of the vehicle in response to the changing context. Distraction from this task, however, is not so easily defined. As Lee et al. note [56], the definition of distraction in the literature is varied and ranges from reasons to outcomes, yet one common theme is that of attention and its distribution between multiple ongoing tasks [27]. Here we adopt their common definition:

> "Driver distraction is a diversion of attention away from activities critical for safe driving towards a competing activity." [56]

At a most basic level, the integration of CUIs with smartphones has allowed these devices to afford a varied and complex set of tasks ideally without demanding the visual or physical attention of the driver allowing the driver to stay attentive and minimise delays to the driver's reactions [47, 63]. These are commonly employed for placing calls, destination entry [67], controlling entertainment [33], reading and transcribing messages, or similar tasks. It has been shown that voice command based systems are less demanding, and therefor presumably safer, than visual-manual interfaces [67].

However, this may not be as straightforward as it seems. There are studies from the driving domain that argue that complex voice interactions show significant attention demands [67]. Using speech-based interfaces while driving can be cognitively captivating, which has the possibility to significantly impair the driving task [38]. It is therefore important to understand the demand that such interfaces may place on drivers, and understand if it is a matter of CUI design

or the modality itself that increases the driver's perceived cognitive load, and with it the vulnerability for distraction [90].

This research can not only provide important insights for the development of CUIs aimed at drivers and passengers in motor vehicles, but presents an opportunity to inform the design of CUIs more widely. Understanding what about a certain conversational interface demands more attention than others while driving is an important step to providing guidelines on designing CUIs that are easy to use across all domains of use which are expected to be concurrent with ongoing activities.

The context of driving, both inside and outside of the car is an important aspect of distraction that is often overlooked in the literature on the subject. While mobile phone use is certainly a distraction, other aspects of the context can also distract the driver and it is in these complex situations that CUIs may provide some reduction in distraction, and possibly therefor in accidents. While there is some evidence that people use their mobile devices less when driving with children in the car [81], having children in the car is also a major source of distraction. Rudin-Brown et al. [83] found that their participants interacted with children in the back seat for 'potentially distracting activities' 12 times more than they interacted with their mobile phone. We have seen one attempt at designing a CUI to reduce this, Gordon et al. [33] designed a conversational agent as something of a 'go-between' in the interactions between children and adults over media use in the car. This is an interesting first step, however a more complex agent could provide better results.

Tchankue et al. [90] propose complementing the dialogue model with a sensor based context-aware module to tailor the interaction to the distraction and driving context. While this is still marked for 'future work', the possibility to time-shift interactions away from moments when the driver is engaged in a safety critical driving task by using conversational strategies to delay their interaction with the system, or to stall their interaction with others in the car or over real-time communication links who are making demands on their attention at that time could be highly beneficial. A simple 'barge-in' by the agent demanding a pause in the communications would be one crude first step, but a more subtle and complex management of the communications between drivers and others are not outwith the realms of possibilities. Of course, the problem that presents itself here is that of agency – one that is common across all contexts – in that any system that overrides the wishes of the user (here by making their communications slower, or stopping them for a time altogether) must negotiate that control effectively and transparently or run the risk of being turned off.

4.4 Games

There were a number of games and interfaces with game elements presented in the papers described here, while the act of motivating CUI interaction through the use of game elements (such as the probe described in [18]) is a long standing technique in HCI and related fields, what is more interesting are the ways in which ludic principles are applied to to CUIs, and how games themselves can be improved with the inclusion and development of CUI elements.

CUIs by their very nature lend themselves to some aspects of ludic design [32] and narrative engagement [24]. As a 'game board' CUIs have been shown to be used for people to play together [6, 73] on applications that provide quiz-like interactions, yet to some extent the repeatedly presented limitation of their opaque affordances can be seen as an opportunity for ludic interaction:

> "In every situation, the ludic self is in search of new possibilities in order to increase the field of possible action." [24]

Choose-your-own-adventure and story-based games [6] are also implemented and used in CUIs, providing the user the opportunity to respond to questions from the agent to advance the story. These games are interesting in that they, primarily, take advantage of the stock text-to-speech modules included on commercial CUIs resulting in "a monotone Alexa saying things like, 'Oh, my, I'm really scared now,' in the exact same tone she replies that she's turned your lights off" [12]. Providing tools to allow developers and designers to imbue emotion in the generated speech, taking advantage of the research done for assistive technologies such as [31] for example, would allow more expressive and immersive games without the expense of voice actors. This would also give the possibility to procedurally generate the game or story dialogue. Combining this with the affordances of mobile devices offers the possibility to expand and deepen the experiences currently on offer with games such as 'Cargo' [66]. This pervasive multiplayer mixed-reality game involved a team is trying to help one member to escape from the city before being caught by the police. A software agent calls the player to help them, a game mechanic that could be incorporated with any number of narrative driven games. By building upon mechanics of vocal analysis such as those used in the Karaoke game SLIONS [68] or the various speech therapy tools [26, 36] (which naturally lend themselves to personalised models on a personal device) in combination with the more complex text-to-speech generation provides opportunities for more complex, deeper, and emotive interactions through CUIs motivated by gaming principles and narratives.

5 DISCUSSION

Beyond the themes and domains, we would like to draw attention to three overarching areas in the adoption and use of CUIs on mobile devices. The first centres around the opportunities and challenges of *multi modal CUIs*, the second on the impact that mobility can play in speech *breakdown and recovery*, and finally the opportunities that the mobile device's access to the personal *context* of the user afford.

5.1 Multi-Modality in CUIs

An important design consideration in CUIs for mobile devices is, of course, the rest of the mobile device beyond the microphone and speaker typically associated with the audio communication channel. In the papers discussed here, it has been shown that combining a CUI with a touchscreen for either input or output can have many advantages. There was a preference to read output and speak input [85], harnessing the information density of written language in comparison to Text To Speech (TTS) output, and the ease of speaking in comparison to using a touch screen keyboard.

Beyond replacing parts of the CUI with other modalities, implementing complimentary interface components has been shown to

help with accuracy [71], or to provide contextual [68, 96] and historical [43] information on the ongoing interaction. Coupled interaction across modalities therefore provides opportunities for a number of traditional challenges in CUIs, which may not always have to be addressed exclusively through audio. Explicitly *De-coupling* the CUI from the rest of the interaction can also be explored to take advantage of the unused modality for complimentary, but distinct tasks. For example a speech therapy task can be lightly coupled to a platform game [36], and furthering interaction research on this path could involve the use of techniques for pro-active CUIs [61] combined with greater awareness of context and ongoing use to initiate parallel activities of the user's choosing (for example language learning) when the opportunity presents itself.

5.2 Breakdowns and Recovery

People regularly fail to interact with speech interfaces [18], and detecting and recovering from communication breakdowns remains a key challenge [7, 18].

Repairing misunderstandings between humans and machines is a fundamental part of human-computer interaction, mechanisms are needed to recover from the inevitable moment when the human and the computer fail to understand each other [18, 89]. One of the long standing goals of natural language processing is to allow conversation repair to be part of human machine interaction [37]. What is of interest in the work we cover is that it provides opportunities for the mechanisms employed in speech therapy and language learning to be applied to CUIs in general. Applying phonological hints [13, 59] (drawing out and emphasising the phonemes in misunderstood words on reply) or communication scaffolding [58, 78] (replying with a mirror of a successful interaction) to aid reformulation of unsuccessful queries for detectable misinterpretations by the CUI are one approach. Greater ongoing awareness of the context of use would provide opportunities to address issues perceived as failure of attention on the part of the CUI, which would need to be communicated both explicitly through social dialog and implicitly using back-channels on other modalities.

Some of the observed breakdowns were due to the range of contexts and participants that developing on a mobile device affords. One aspect of that is technical, with different microphones on different devices potentially causing issues of accuracy [72, 94], and challenges of varying connectivity and processing power [43].

The highly variable social and personal contexts of the users also raises issues. Socially responsive interfaces react to how people interact with them [16], lexically [57] or non-verbally [61], yet the range of adaption afforded is currently limited. The diversity of languages, vernaculars, dialects, and people understood and supported by CUIs is an important, yet incredibly difficult challenge. This shouldn't be seen just as a problem of improving speech-to-text accuracy for specific populations (for example, adapting to the slower speech and inter-syllabic silence of elderly users [53]), but more widely on the understanding and adapting to *how* different groups of people speak; their idioms, tropes, and methods for imbuing emotional and social subtlety in language.

5.3 Context

The most interesting and important thing that defines the opportunities for CUIs on mobile devices is their almost unparalleled access to the context of the user. Understanding the ongoing activity of the user through the sensors on mobile devices has been a increasingly focused upon area of research in recent years [35, 50, 76, 80, 88], leading to both general and domain specific recognisers, as well as taking advantage of connected devices such as smartwatches to improve the recognition [8]. Indeed, current generations of commercially available smartwatches ship with limited activity recognition for exercise [5], falls [4], and heart problems [2]. Beyond physical activity, current smartwatches also provide very limited monitoring of noise levels, providing alerts when in situations that could damage hearing [3]. This continuous monitoring of the noise level can be seen as a precursor to more complex monitoring of the audio channel to understand the user's context. McMillan et al. [64] suggested a number of personal and shared opportunities that listening to the ongoing conversations of users could afford. Even simple noise level monitoring could be used to improve CUI interactions, adjusting the output in volume, speed, and inter-syllable spacing can improve understandability at high noise levels. The CUI could also adjust the output and the input around regular noises, or simply suggest that the user move to an area with better acoustic properties rather than mis-recognising their speech. Such adjustments could also be developed around different types of detected activity, allowing for interrupted input and timing responses to be better in keeping with the ongoing action and the temporally shifting amount of attention the user can spare through the activity.

With more complex social context understood through listening to the ongoing conversations of the user as outlined in [64] there are more interesting, and technically challenging, opportunities to influence interactions with the CUI. Adjusting the emotional valance of the generated conversation to account for the detected mood of the user could be used to provide an efficiency fit (for example, when they are detected to be addressing others quickly and with brevity then the CUI could trim unnecessary politeness and opportunistic dialogue embellishments) or to provide intervention to alleviate an undesirable change in mood through pro-actively engaging in conversation or adding embellishments to ongoing interactions on the topic. Beyond current emotional state, such a *Qualified Self* detection module [64] in collaboration with a CUI could intervene when interpersonal communications are detected to be either lacking in amount (by providing a quasi conversational partner), stilted (by providing topics for conversion [45]), or confrontational [1].

6 CONCLUSION

In this paper, we have surveyed research that investigates or uses CUIs on mobile devices. By highlighting the specific opportunities afforded by mobile devices for the design and development of CUI interactions, we hope to encourage further work on the integration of context, the inclusion of multiple modalities, and providing more complex opportunities to recover from communication breakdowns with conversational interfaces.

REFERENCES

[1] Salvatore Andolina, Valeria Orso, Hendrik Schneider, Khalil Klouche, Tuukka Ruotsalo, Luciano Gamberini, and Giulio Jacucci. 2018. SearchBot: Supporting Voice Conversations with Proactive Search. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '18)*. Association for Computing Machinery, Jersey City, NJ, USA, 9–12. https://doi.org/10.1145/3272973.3272990

[2] Apple. 2020. Heart Health Notifications on Your Apple Watch. https://support.apple.com/en-us/HT208931.

[3] Apple. 2020. Measure Noise Levels with Apple Watch. https://support.apple.com/en-gb/guide/watch/apd00a43a9cb/watchos.

[4] Apple. 2020. Use Fall Detection with Apple Watch. https://support.apple.com/en-us/HT208944.

[5] Apple. 2020. Workout Types on Apple Watch. https://support.apple.com/en-mide/HT207934.

[6] Diana Beirl, Nicola Yuill, and Yvonne Rogers. 2019. Using Voice Assistant Skills in Family Life. In *A Wide Lens: Combining Embodied, Enactive, Extended, and Embedded Learning in Collaborative Settings*, Vol. 1. International Society of the Learning Sciences, 96–103.

[7] Erin Beneteau, Olivia K. Richards, Mingrui Zhang, Julie A. Kientz, Jason Yip, and Alexis Hiniker. 2019. Communication Breakdowns Between Families and Alexa. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*. ACM Press, Glasgow, Scotland Uk, 1–13. https://doi.org/10.1145/3290605.3300473

[8] Sourav Bhattacharya and Nicholas D. Lane. 2016. From Smart to Deep: Robust Activity Recognition on Smartwatches Using Deep Learning. In *2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*. 1–6. https://doi.org/10.1109/PERCOMW.2016.7457169

[9] Bhakti Bhikne, Anirudha Joshi, Manjiri Joshi, Shashank Ahire, and Nimish Maravi. 2018. How Much Faster Can You Type by Speaking in Hindi?: Comparing Keyboard-Only and Keyboard+Speech Text Entry. In *Proceedings of the 9th International Conference on HCI IndiaHCI 2018 - IndiaHCI 18*. ACM Press, Bangalore, India, 20–28. https://doi.org/10.1145/3297121.3297123

[10] Timothy Bickmore, Ha Trinh, Reza Asadi, and Stefan Olafsson. 2018. Safety First: Conversational Agents for Health Care. In *Studies in Conversational UX Design*, Robert J. Moore, Margaret H. Szymanski, Raphael Arar, and Guang-Jie Ren (Eds.). Springer International Publishing, Cham, 33–57. https://doi.org/10.1007/978-3-319-95579-7_3

[11] Kirsten Boehner, Janet Vertesi, Phoebe Sengers, and Paul Dourish. 2007. How HCI Interprets the Probes. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. Association for Computing Machinery, San Jose, California, USA, 1077–1086. https://doi.org/10.1145/1240624.1240789

[12] Alina Bradford and Jenny McGrath. 2019. The Best Games to Play With Alexa. https://www.digitaltrends.com/home/best-games-to-play-with-alexa/.

[13] Catherine P. Browman and Louis Goldstein. 1989. Articulatory Gestures as Phonological Units. *Phonology* 6, 2 (Aug. 1989), 201–251. https://doi.org/10.1017/S0952675700001019

[14] Barry Brown, Moira McGregor, and Donald McMillan. 2014. 100 Days of iPhone Use: Understanding the Details of Mobile Device Use. In *Proceedings of the 16th International Conference on Human-Computer Interaction with Mobile Devices & Services - MobileHCI '14*. ACM Press, Toronto, ON, Canada, 223–232. https://doi.org/10.1145/2628363.2628377

[15] Barry Brown, Moira McGregor, and Donald McMillan. 2015. Searchable Objects: Search in Everyday Conversation. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. ACM, Vancouver, BC, Canada, 508–517. https://doi.org/10.1145/2675133.2675206

[16] Charles Callaway and Khalil Sima'an. 2006. Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship. *Computational Linguistics* 32, 3 (Sept. 2006), 451–452. https://doi.org/10.1162/coli.2006.32.3.451

[17] Yung Fu Chang, C.S. Chen, and Hao Zhou. 2009. Smart Phone for Mobile Commerce. *Computer Standards & Interfaces* 31, 4 (June 2009), 740–747. https://doi.org/10.1016/j.csi.2008.09.016

[18] Yi Cheng, Kate Yen, Yeqi Chen, Sijin Chen, and Alexis Hiniker. 2018. Why Doesn't It Work?: Voice-Driven Interfaces and Young Children's Communication Repair Strategies. In *Proceedings of the 17th ACM Conference on Interaction Design and Children - IDC '18*. ACM Press, Trondheim, Norway, 337–348. https://doi.org/10.1145/3202185.3202749

[19] Leigh Clark, Phillip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew Aylett, João Cabral, Cosmin Munteanu, and Benjamin Cowan. 2019. The State of Speech in HCI: Trends, Themes and Challenges. *Interacting with Computers* (Sept. 2019), iwz016. https://doi.org/10.1093/iwc/iwz016 arXiv:1810.06828

[20] G. Clough, A. C. Jones, P. McAndrew, and E. Scanlon. 2008. Informal Learning with PDAs and Smartphones. *Journal of Computer Assisted Learning* 24, 5 (2008), 359–371. https://doi.org/10.1111/j.1365-2729.2007.00268.x

[21] D. Coniam. 1999. Voice Recognition Software Accuracy with Second Language Speakers of English. *System* 27, 1 (March 1999), 49–64. https://doi.org/10.1016/S0346-251X(98)00049-9

[22] Sebastien Cuendet, Indrani Medhi, Kalika Bali, and Edward Cutrell. 2013. VideoKheti: Making Video Content Accessible to Low-Literate and Novice Users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*. ACM Press, Paris, France, 2833. https://doi.org/10.1145/2470654.2481392

[23] D. Damos. 1991. *Multiple Task Performance*. CRC Press.

[24] Jos de Mul. 2015. The Game of Life: Narrative and Ludic Identity Formation in Computer Games. In *Representations of Internarrative Identity*, Lori Way (Ed.). Palgrave Macmillan UK, London, 159–187. https://doi.org/10.1057/9781137462534_10

[25] E. Ray Dorsey and Eric J. Topol. 2016. State of Telehealth. *New England Journal of Medicine* 375, 2 (July 2016), 154–161. https://doi.org/10.1056/NEJMra1601705

[26] Jared Duval, Zachary Rubin, Elena Márquez Segura, Natalie Friedman, Milla Zlatanov, Louise Yang, and Sri Kurniawan. 2018. SpokeIt: Building a Mobile Speech Therapy Experience. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services - MobileHCI '18*. ACM Press, Barcelona, Spain, 1–12. https://doi.org/10.1145/3229434.3229484

[27] Justin Edwards, He Liu, Tianyu Zhou, Sandy J. J. Gould, Leigh Clark, Philip Doyle, and Benjamin R. Cowan. 2019. Multitasking with Alexa: How Using Intelligent Personal Assistants Impacts Language-Based Primary Task Performance. In *Proceedings of the 1st International Conference on Conversational User Interfaces - CUI '19*. ACM Press, Dublin, Ireland, 1–7. https://doi.org/10.1145/3342775.3342785

[28] Yasmine N. El-Glaly and Francis Quek. 2014. Digital Reading Support for The Blind by Multimodal Interaction. In *Proceedings of the 16th International Conference on Multimodal Interaction - ICMI '14*. ACM Press, Istanbul, Turkey, 439–446. https://doi.org/10.1145/2663204.2663266

[29] Maxine Eskenazi. 1999. Using Automatic Speech Processing for Foreign Language Pronunciation Tutoring: Some Issues and a Prototype. *Language Learning & Technology* 2, 2 (1999), 13.

[30] Ahmed Fadhil. 2018. Beyond Patient Monitoring: Conversational Agents Role in Telemedicine & Healthcare Support For Home-Living Elderly Individuals. *arXiv:1803.06000 [cs]* (March 2018). arXiv:cs/1803.06000

[31] Alexander J. Fiannaca, Ann Paradiso, Jon Campbell, and Meredith Ringel Morris. 2018. Voicesetting: Voice Authoring UIs for Improved Expressivity in Augmentative Communication. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*. ACM Press, Montreal QC, Canada, 1–12. https://doi.org/10.1145/3173574.3173857

[32] William W. Gaver, John Bowers, Andrew Boucher, Hans Gellerson, Sarah Pennington, Albrecht Schmidt, Anthony Steed, Nicholas Villars, and Brendan Walker. 2004. The Drift Table: Designing for Ludic Engagement. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems (CHI EA '04)*. Association for Computing Machinery, Vienna, Austria, 885–900. https://doi.org/10.1145/985921.985947

[33] Michal Gordon and Cynthia Breazeal. 2015. Designing a Virtual Assistant for In-Car Child Entertainment. In *Proceedings of the 14th International Conference on Interaction Design and Children - IDC '15*. ACM Press, Boston, Massachusetts, 359–362. https://doi.org/10.1145/2771839.2771916

[34] Connor Graham, Mark Rouncefield, Martin Gibbs, Frank Vetere, and Keith Cheverst. 2007. How Probes Work. In *Proceedings of the 19th Australasian Conference on Computer-Human Interaction: Entertaining User Interfaces (OZCHI '07)*. Association for Computing Machinery, Adelaide, Australia, 29–37. https://doi.org/10.1145/1324892.1324899

[35] Norbert Győrbíró, Ákos Fábián, and Gergely Hományi. 2009. An Activity Recognition System For Mobile Phones. *Mobile Networks and Applications* 14, 1 (Feb. 2009), 82–91. https://doi.org/10.1007/s11036-008-0112-y

[36] Adam Hair, Penelope Monroe, Beena Ahmed, Kirrie J. Ballard, and Ricardo Gutierrez-Osuna. 2018. Apraxia World: A Speech Therapy Game for Children with Speech Sound Disorders. In *Proceedings of the 17th ACM Conference on Interaction Design and Children - IDC '18*. ACM Press, Trondheim, Norway, 119–131. https://doi.org/10.1145/3202185.3202733

[37] Philip J. Hayes and D. Raj Reddy. 1983. Steps toward Graceful Interaction in Spoken and Written Man-Machine Communication. *International Journal of Man-Machine Studies* 19, 3 (Sept. 1983), 231–284. https://doi.org/10.1016/S0020-7373(83)80049-2

[38] Jibo He, Alex Chaparro, Bobby Nguyen, Rondell Burge, Joseph Crandall, Barbara Chaparro, Rui Ni, and Shi Cao. 2013. Texting While Driving: Is Speech-Based Texting Less Risky than Handheld Texting?. In *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '13*. ACM Press, Eindhoven, Netherlands, 124–130. https://doi.org/10.1145/2516540.2516560

[39] Rebecca Hincks. 2003. *Speech Technologies for Pronunciation Feedback and Evaluation*.

[40] Adam O. Horvath, A. C. Del Re, Christoph Flückiger, and Dianne Symonds. 2011. Alliance in Individual Psychotherapy. *Psychotherapy* 48, 1 (2011), 9–16. https://doi.org/10.1037/a0022186

[41] David Huggins-Daines, Mohit Kumar, Arthur Chan, Alan W. Black, Mosur Ravishankar, and Alexander I. Rudnicky. 2006. Pocketsphinx: A Free, Real-Time Continuous Speech Recognition System for Hand-Held Devices. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, Vol. 1. IEEE, I–I.

[42] Sami Hulkko, Tuuli Mattelmäki, Katja Virtanen, and Turkka Keinonen. 2004. Mobile Probes. In *Proceedings of the Third Nordic Conference on Human-Computer Interaction (NordiCHI '04)*. Association for Computing Machinery, Tampere, Finland, 43–51. https://doi.org/10.1145/1028014.1028020

[43] Mohit Jain, Pratyush Kumar, Ishita Bhansali, Q. Vera Liao, Khai Truong, and Shwetak Patel. 2018. FarmChat: A Conversational Agent to Answer Farmer Queries. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (Dec. 2018), 1–22. https://doi.org/10.1145/3287048

[44] Mohit Jain, Pratyush Kumar, Ramachandra Kota, and Shwetak N. Patel. 2018. Evaluating and Informing the Design of Chatbots. In *Proceedings of the 2018 on Designing Interactive Systems Conference 2018 - DIS '18*. ACM Press, Hong Kong, China, 895–906. https://doi.org/10.1145/3196709.3196735

[45] Pradthana Jarusriboonchai, Thomas Olsson, and Kaisa Väänänen-Vainio-Mattila. 2014. User Experience of Proactive Audio-Based Social Devices: A Wizard-of-Oz Study. In *Proceedings of the 13th International Conference on Mobile and Ubiquitous Multimedia - MUM '14*. ACM Press, Melbourne, Victoria, Australia, 98–106. https://doi.org/10.1145/2677972.2677995

[46] Scott D. Johnson, Steven R. Aragon, and Najmuddin Shaik. 2000. Comparative Analysis of Learner Satisfaction and Learning Outcomes in Online and Face-to-Face Learning Environments. *Journal of Interactive Learning Research* 11, 1 (2000), 29–49.

[47] Marcel Adam Just, Timothy A. Keller, and Jacquelyn Cynkar. 2008. A Decrease in Brain Activation Associated with Driving When Listening to Someone Speak. *Brain Research* 1205 (April 2008), 70–80. https://doi.org/10.1016/j.brainres.2007.12.075

[48] Sin-Hwa Kang, Andrew W. Feng, Anton Leuski, Dan Casas, and Ari Shapiro. 2015. The Effect of An Animated Virtual Character on Mobile Chat Interactions. In *Proceedings of the 3rd International Conference on Human-Agent Interaction (HAI '15)*. Association for Computing Machinery, Daegu, Kyungpook, Republic of Korea, 105–112. https://doi.org/10.1145/2814940.2814957

[49] Sin-Hwa Kang, James H. Watt, and Sasi Kanth Ala. 2008. Social Copresence in Anonymous Social Interactions Using a Mobile Video Telephone. In *Proceeding of the Twenty-Sixth Annual CHI Conference on Human Factors in Computing Systems - CHI '08*. ACM Press, Florence, Italy, 1535. https://doi.org/10.1145/1357054.1357295

[50] Wazir Zada Khan, Yang Xiang, Mohammed Y Aalsalem, and Quratulain Arshad. First 2013. Mobile Phone Sensing Systems: A Survey. *IEEE Communications Surveys Tutorials* 15, 1 (First 2013), 402–427. https://doi.org/10.1109/SURV.2012.031412.00077

[51] Bret Kinsella and Ava Mutchler. 2019. *U.S. Smart Speaker Consumer Adoption Report 2019.* Technical Report. Voicebot.ai and Voicify.

[52] Anuj Kumar, Pooja Reddy, Anuj Tewari, Rajat Agrawal, and Matthew Kam. 2012. Improving Literacy in Developing Countries Using Speech Recognition-Supported Games on Mobile Devices. In *Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems - CHI '12*. ACM Press, Austin, Texas, USA, 1149. https://doi.org/10.1145/2207676.2208564

[53] Soonil Kwon, Sung-Jae Kim, and Joon Yeon Choeh. 2016. Preprocessing for Elderly Speech Recognition of Smart Devices. *Computer Speech & Language* 36 (March 2016), 110–121. https://doi.org/10.1016/j.csl.2015.09.002

[54] Liliana Laranjo, Adam G Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie Y S Lau, and Enrico Coiera. 2018. Conversational Agents in Healthcare: A Systematic Review. *Journal of the American Medical Informatics Association* 25, 9 (Sept. 2018), 1248–1258. https://doi.org/10.1093/jamia/ocy072

[55] David R. Large, Gary Burnett, Ben Anyasodo, and Lee Skrypchuk. 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 - Automotive'UI 16*. ACM Press, Ann Arbor, MI, USA, 67–74. https://doi.org/10.1145/3003715.3005408

[56] John D. Lee, Kristie L. Young, and Michael A. Regan. 2008. Defining Driver Distraction. In *Driver Distraction: Theory, Effects, and Mitigation*. CRC Press, 31–40.

[57] Min Kyung Lee, Jodi Forlizzi, Sara Kiesler, Paul Rybski, John Antanitis, and Sarun Savetsila. 2012. Personalization in HRI: A Longitudinal Field Experiment. In *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 319–326.

[58] Nicole Liboiron and Gloria Soto. 2006. Shared Storybook Reading with a Student Who Uses Alternative and Augmentative Communication: A Description of Scaffolding Practices. *Child Language Teaching and Therapy* 22, 1 (Feb. 2006), 69–95. https://doi.org/10.1191/0265659006ct298oa

[59] Anders Löfqvist. 1990. Speech as Audible Gestures. In *Speech Production and Speech Modelling*, William J. Hardcastle and Alain Marchal (Eds.). Springer Netherlands, Dordrecht, 289–322. https://doi.org/10.1007/978-94-009-2037-8_12

[60] Hong Lu, Wei Pan, Nicholas D. Lane, Tanzeem Choudhury, and Andrew T. Campbell. 2009. SoundSense: Scalable Sound Sensing for People-Centric Applications on Mobile Phones. In *Proceedings of the 7th International Conference on Mobile Systems, Applications, and Services - Mobisys '09*. ACM Press, Wroclaw, Poland, 165. https://doi.org/10.1145/1555816.1555834

[61] Nichola Lubold, Erin Walker, and Heather Pon-Barry. 2016. Effects of Voice-Adaptation and Social Dialogue on Perceptions of a Robotic Learning Companion. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, Christchurch, New Zealand, 255–262. https://doi.org/10.1109/HRI.2016.7451760

[62] Moira McGregor and John C. Tang. 2017. More to Meetings: Challenges in Using Speech-Based Technology to Support Meetings. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, Portland, Oregon, USA, 2208–2220. https://doi.org/10.1145/2998181.2998335

[63] Joshua D. McKeever, Maria T. Schultheis, Vennila Padmanaban, and Allison Blasco. 2013. Driver Performance While Texting: Even a Little Is Too Much. *Traffic Injury Prevention* 14, 2 (Jan. 2013), 132–137. https://doi.org/10.1080/15389588.2012.699695

[64] Donald McMillan, Antoine Loriette, and Barry Brown. 2015. Repurposing Conversation: Experiments with the Continuous Speech Stream. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 3953–3962. https://doi.org/10.1145/2702123.2702532

[65] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]* (Sept. 2013). arXiv:cs/1301.3781

[66] Stuart Moran, Nadia Pantidi, Khaled Bachour, Joel E. Fischer, Martin Flintham, Tom Rodden, Simon Evans, and Simon Johnson. 2013. Team Reactions to Voiced Agent Instructions in a Pervasive Game. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces - IUI '13*. ACM Press, Santa Monica, California, USA, 371. https://doi.org/10.1145/2449396.2449445

[67] Daniel Munger, Bruce Mehler, Bryan Reimer, Jonathan Dobres, Anthony Pettinato, Brahmi Pugh, and Joseph F. Coughlin. 2014. A Simulation Study Examining Smartphone Destination Entry While Driving. In *Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '14*. ACM Press, Seattle, WA, USA, 1–5. https://doi.org/10.1145/2667317.2667349

[68] Dania Murad, Riwu Wang, Douglas Turnbull, and Ye Wang. 2018. SLIONS: A Karaoke Application to Enhance Foreign Language Learning. In *2018 ACM Multimedia Conference on Multimedia Conference - MM '18*. ACM Press, Seoul, Republic of Korea, 1679–1687. https://doi.org/10.1145/3240508.3240691

[69] Kyo-Joong Oh, Dongkun Lee, Byungsoo Ko, and Ho-Jin Choi. 2017. A Chatbot for Psychiatric Counseling in Mental Healthcare Service Based on Emotional Dialogue Analysis and Sentence Generation. In *2017 18th IEEE International Conference on Mobile Data Management (MDM)*. 371–375. https://doi.org/10.1109/MDM.2017.64

[70] Sun Owen, Chen Jessica, and Magrabi Farah. 2018. Using Voice-Activated Conversational Interfaces for Reporting Patient Safety Incidents: A Technical Feasibility and Pilot Usability Study. *Studies in Health Technology and Informatics* (2018), 139–144. https://doi.org/10.3233/978-1-61499-890-7-139

[71] Shun Ozaki, Takuo Matsunobe, Takashi Yoshino, and Aguri Shigeno. 2011. Design of a Face-to-Face Multilingual Communication System for a Handheld Device in the Medical Field. In *Human-Computer Interaction. Interaction Techniques and Environments*, Julie A. Jacko (Ed.). Vol. 6762. Springer Berlin Heidelberg, Berlin, Heidelberg, 378–386. https://doi.org/10.1007/978-3-642-21605-3_42

[72] Avinash Parnandi, Virendra Karappa, Tian Lan, Mostafa Shahin, Jacqueline McKechnie, Kirrie Ballard, Beena Ahmed, and Ricardo Gutierrez-Osuna. 2015. Development of a Remote Therapy Tool for Childhood Apraxia of Speech. *ACM Transactions on Accessible Computing* 7, 3 (Nov. 2015), 1–23. https://doi.org/10.1145/2776895

[73] Martin Porcheron, Joel E. Fischer, Moira McGregor, Barry Brown, Ewa Luger, Heloisa Candello, and Kenton O'Hara. 2017. Talking with Conversational Agents in Collaborative Action. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17 Companion)*. ACM, Portland, Oregon, USA, 431–436. https://doi.org/10.1145/3022198.3022666

[74] Martin Porcheron, Joel E. Fischer, and Sarah Sharples. 2017. "Do Animals Have Accents?": Talking with Agents in Multi-Party Conversation. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, Portland, Oregon, USA, 207–219. https://doi.org/10.1145/2998181.2998298

[75] Fang Qiao, Jahanzeb Sherwani, and Roni Rosenfeld. 2010. Small-Vocabulary Speech Recognition for Resource-Scarce Languages. In *Proceedings of the First ACM Symposium on Computing for Development - ACM DEV '10*. ACM Press, London, United Kingdom, 1. https://doi.org/10.1145/1926180.1926184

[76] Sreenivasan Ramasamy Ramamurthy and Nirmalya Roy. 2018. Recent Trends in Machine Learning for Human Activity Recognition—A Survey. *WIREs Data Mining and Knowledge Discovery* 8, 4 (July 2018), e1254. https://doi.org/10.1002/widm.1254

[77] Agha Ali Raza, Mansoor Pervaiz, Christina Milo, Samia Razaq, Guy Alster, Jahanzeb Sherwani, Umar Saif, and Roni Rosenfeld. 2012. Viral Entertainment as a Vehicle for Disseminating Speech-Based Services to Low-Literate Users. In *Proceedings of the Fifth International Conference on Information and Communication Technologies and Development - ICTD '12*. ACM Press, Atlanta, Georgia, 350. https://doi.org/10.1145/2160673.2160715

[78] Jane Remington-Gurney. 2013. Scaffolding Conversations Using Augmentative and Communication (AAC) Alternative. (2013), 25.

[79] Nikki Rickard, Hussain-Abdulah Arjmand, David Bakker, and Elizabeth Seabrook. 2016. Development of a Mobile Phone App to Support Self-Monitoring of Emotional Well-Being: A Mental Health Digital Innovation. *JMIR Mental Health* 3, 4 (Nov. 2016), e49. https://doi.org/10.2196/mental.6202

[80] Charissa Ann Ronao and Sung-Bae Cho. 2016. Human Activity Recognition with Smartphone Sensors Using Deep Learning Neural Networks. *Expert Systems with Applications* 59 (Oct. 2016), 235–244. https://doi.org/10.1016/j.eswa.2016.04.032

[81] Linda Roney, Pina Violano, Greg Klaus, Rebecca Lofthouse, and James Dziura. 2013. Distracted Driving Behaviors of Adults While Children Are in the Car. *Journal of trauma and acute care surgery* 75, 4 (2013), S290–S295.

[82] Sherry Ruan, Jacob O. Wobbrock, Kenny Liou, Andrew Ng, and James A. Landay. 2018. Comparing Speech and Keyboard Text Entry for Short Messages in Two Languages on Touchscreen Phones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (Jan. 2018), 1–23. https://doi.org/10.1145/3161187

[83] Christina M. Rudin-Brown, S. Koppel, Belinda Clark, and J. Charlton. 2012. Prevalence of Mobile Phone vs. Child-Related Driver Distraction in a Sample of Families with Young Children. *Journal of the Australasian College of Road Safety* 23, 2 (2012), 58.

[84] Robert P. Schumaker and Hsinchun Chen. 2010. Interaction Analysis of the ALICE Chatterbot: A Two-Study Investigation of Dialog and Domain Questioning. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 40, 1 (Jan. 2010), 40–51. https://doi.org/10.1109/TSMCA.2009.2029603

[85] Sajad Shirali-Shahreza, Gerald Penn, Ravin Balakrishnan, and Yashar Ganjali. 2013. SeeSay and HearSay CAPTCHA for Mobile Interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*. ACM Press, Paris, France, 2147. https://doi.org/10.1145/2470654.2481295

[86] Laura Silver and Stefan Cornibert. 2019. *Smartphone Ownership Is Growing Rapidly Around the World, but Not Always Equally*. Technical Report. Pew Research Center.

[87] Donggil Song, Eun Young Oh, and Marilyn Rice. 2017. Interacting with a Conversational Agent System for Educational Purposes in Online Courses. In *2017 10th International Conference on Human System Interactions (HSI)*. 78–82. https://doi.org/10.1109/HSI.2017.8005002

[88] Xing Su, Hanghang Tong, and Ping Ji. 2014. Activity Recognition with Smartphone Sensors. *Tsinghua Science and Technology* 19, 3 (June 2014), 235–249. https://doi.org/10.1109/TST.2014.6838194

[89] Lucy Suchman. 1987. *Plans and Situated Actions: The Problem of Human-Machine Communication*. Cambridge University Press.

[90] Patrick Tchankue, Janet Wesson, and Dieter Vogts. 2012. Are Mobile In-Car Communication Systems Feasible? A Usability Study. In *Proceedings of the South African Institute for Computer Scientists and Information Technologists Conference (SAICSIT '12)*. Association for Computing Machinery, Pretoria, South Africa, 262–269. https://doi.org/10.1145/2389836.2389867

[91] Reed V. Tuckson, Margo Edmunds, and Michael L. Hodgkins. 2017. Telehealth. *New England Journal of Medicine* 377, 16 (Oct. 2017), 1585–1592. https://doi.org/10.1056/NEJMsr1503323

[92] Ruolin Wang, Chun Yu, Xing-Dong Yang, Weijie He, and Yuanchun Shi. 2019. EarTouch: Facilitating Smartphone Use for Visually Impaired People in Mobile and Public Scenarios. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*. ACM Press, Glasgow, Scotland Uk, 1–13. https://doi.org/10.1145/3290605.3300254

[93] Yang Wang, Tao Mei, Jingdong Wang, Houqiang Li, and Shipeng Li. 2011. JIGSAW: Interactive Mobile Visual Search with Multimodal Queries. In *Proceedings of the 19th ACM International Conference on Multimedia - MM '11*. ACM Press, Scottsdale, Arizona, USA, 73. https://doi.org/10.1145/2072298.2072310

[94] Cara Wilson, Margot Brereton, Bernd Ploderer, and Laurianne Sitbon. 2018. MyWord: Enhancing Engagement, Interaction and Self-Expression with Minimally-Verbal Children on the Autism Spectrum through a Personal Audio-Visual Dictionary. In *Proceedings of the 17th ACM Conference on Interaction Design and Children - IDC '18*. ACM Press, Trondheim, Norway, 106–118. https://doi.org/10.1145/3202185.3202755

[95] Ye Xu, Mu Lin, Hong Lu, Giuseppe Cardone, Nicholas Lane, Zhenyu Chen, Andrew Campbell, and Tanzeem Choudhury. 2013. Preference, Context and Communities: A Multi-Faceted Approach to Predicting Smartphone App Usage Patterns. (2013), 8.

[96] Daisuke Yamamoto, Keiichiro Oura, Ryota Nishimura, Takahiro Uchiya, Akinobu Lee, Ichi Takumi, and Keiichi Tokuda. 2014. Voice Interaction System with 3D-CG Virtual Agent for Stand-Alone Smartphones. In *Proceedings of the Second International Conference on Human-Agent Interaction - HAI '14*. ACM Press, Tsukuba, Japan, 323–330. https://doi.org/10.1145/2658861.2658874