"Hi! I am the Crowd Tasker" Crowdsourcing through Digital Voice Assistants

Danula Hettiachchi[1], **Zhanna Sarsenbayeva**[1], **Fraser Allison**[1], **Niels van Berkel**[2], **Tilman Dingler**[1], **Gabriele Marini**[1], **Vassilis Kostakos**[1], **Jorge Goncalves**[1]
[1]The University of Melbourne, Melbourne, Australia
[2]University College London, London, United Kingdom
[1]first.last@unimelb.edu.au, [2]n.vanberkel@ucl.ac.uk

ABSTRACT
Inspired by the increasing prevalence of digital voice assistants, we demonstrate the feasibility of using voice interfaces to deploy and complete crowd tasks. We have developed *Crowd Tasker*, a novel system that delivers crowd tasks through a digital voice assistant. In a lab study, we validate our proof-of-concept and show that crowd task performance through a voice assistant is comparable to that of a web interface for voice-compatible and voice-based crowd tasks for native English speakers. We also report on a field study where participants used our system in their homes. We find that crowdsourcing through voice can provide greater flexibility to crowd workers by allowing them to work in brief sessions, enabling multi-tasking, and reducing the time and effort required to initiate tasks. We conclude by proposing a set of design guidelines for the creation of crowd tasks for voice and the development of future voice-based crowdsourcing systems.

Author Keywords
crowdsourcing; smart speakers; digital voice assistants; voice user interface

CCS Concepts
•**Human-centered computing** → **Interaction devices;**
•**Information systems** → **Crowdsourcing;**

INTRODUCTION
Despite the growing popularity of digital voice assistants (such as Alexa, Siri, Google Assistant, and Cortana), they are predominantly used for low-complexity tasks such as setting timers, playing music, checking the weather or regulating a thermostat [4, 44]. Yet, the increasing sophistication of digital voice assistants enables the possibility that more complex tasks, or even sustained work could be conducted through conversational interfaces. Gartner has predicted that 25% of digital workers will use conversational agents on a daily basis by 2021, and that 25% of employee interactions with business applications will be through voice by 2023[1]. This impending shift towards digital voice assistant-enabled work has the potential to instigate voice-based crowdsourcing as a complementary means to conduct crowd work, rather than a replacement to current approaches (*e.g.,* use of online platforms) [33]. Currently, crowd work is nearly always conducted through a screen-based interface such as a desktop computer or a smartphone, and mostly by workers in their own homes [5]. The hands-free and eyes-free nature of voice interaction could be beneficial to these workers–particularly those that juggle crowd work with other responsibilities at home–by allowing them to complete tasks while doing other things around the home. Also, voice-assistants are a promising way to attract new crowd workers, who are only available to complete small amounts of work at opportune moments.

For example, digital voice assistants can allow users to access crowd work more quickly and conveniently by simply talking to the voice assistant whenever they want to work, rather than having to sit at a desk, log in to a device, launch a browser, and finally select a task [33]. These steps can accumulate a substantial amount of lost time if the user is alternating between work and other activities throughout the day. Furthermore, voice interfaces can make crowd work more accessible to users with vision or motor disabilities that make it difficult for them to engage in screen-based work [62]. On the other hand, not all types of crowd tasks are suited for voice-interaction as they may contain indispensable visual elements or involve complex workflows [18].

While previous research has explored speech transcription through smartphone-based voice input [61, 62], these studies involved ad-hoc systems with a single task. The proposed systems do not provide the capability to browse and launch a wider range of crowdsourcing tasks solely using voice commands. Furthermore, there is no prior work investigating the potential of digital voice assistants or smart speakers for crowd work. To facilitate voice-based crowd work, we developed Crowd Tasker, a novel stand-alone voice crowdsourcing application that can be accessed through any device that supports Google Assistant. To assess whether worker performance using voice input is comparable to a regular web interface,

[1]https://www.gartner.com/en/newsroom/press-releases/2019-01-09-gartner-predicts-25-percent-of-digital-workers-will-u

CHI '20, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6708-0/20/04 ...$15.00.
http://dx.doi.org/10.1145/3313831.3376320

we conduct a lab study with 30 participants. We test two types of crowd tasks: voice-compatible (sentiment analysis, comprehension, and text moderation), and voice-based (audio annotation, speech transcription, and emotion labelling), and find that for most tasks, worker accuracy does not significantly differ between the voice assistant and a regular web interface for native English speakers. Subsequently, we conduct a field deployment with another set of 12 participants who completed tasks using a voice assistant in their homes over the course of one week. The aim of the field deployment was to better understand emergent user behaviour and to assess if data quality suffers when completing crowd tasks through a voice assistant when the user is in a less controlled environment. Our results show that participant contributions were of similar quality to those in the lab study. In addition, participants reported that they initiated tasks at opportune moments and worked in brief sessions, while also multitasking when convenient.

Based on our findings, we propose a set of guidelines for the design of future voice-based crowdsourcing systems as well as best practices for creating voice-compatible crowd tasks.

RELATED WORK

Voice Interaction and Digital Voice Assistants

While voice interaction technologies have been developed for a number of decades, there has been renewed interest in the topic with the popularity and growing availability of digital voice assistants. Recent work by Bentley *et al.* [4] examines the use of digital voice assistants in 88 households. Their speech log analysis reveals that users engage with smart speakers through short sessions throughout the day as opposed to using the device for longer periods of time to complete a series of tasks. Furthermore, they show that users more frequently utilise smart speakers when compared to phone-based voice assistants. They identify Music, Information (*e.g.,* asking for spellings) and Automation (*e.g.,* turning off lights) as most frequently used command categories.

Several studies have compared voice input to manual input for the same task and report that voice input rates well on engagement, but poorly on usability and sense of control [2, 45]. Due to a lack of typical user interaction signals like mouse clicks and scroll movements, measuring and evaluating user satisfaction on voice interfaces greatly differs from traditional screen based interfaces. In a recent survey, Kocaballi *et al.* [39] examine a number of studies that aim to understand and measure user experience in conversational interfaces. For example, Hashemi *et al.* [32] propose to model user satisfaction by creating intent sensitive word embeddings or by representing user interactions as a sequence of user intents. The literature also proposes design guidelines that can create better voice user interfaces [15, 16, 46, 47]. However, research highlights that voice interfaces require better theories and more design guidelines, due to persistent usability issues [14, 48].

Further, research shows that assimilation bias can have an impact on performance in voice user interfaces. In a study where participants were asked to use a voice based calendar application, Myers *et al.* [49] report that participants with increased experience with voice user interfaces took less time with tasks. In addition to the experience, language proficiency is known to impact the usability of digital voice assistants. Pyae *et al.* [52] report that native English speakers had a better overall user experience when compared to non-native English speakers when using Google Home devices. Research has also shown that matching the personality of the voice assistant and the user's expectations can result in higher likeability and trust for assistants [6]. In a study involving older adults, Chattaraman *et al.* [9] report that users' internet competency and the digital assistant's conversational style can have significant interaction effects on social (*e.g.,* trust in the system), functional (*e.g.,* perceived ease of using the system), and behavioural intent outcomes. Several other studies have also confirmed that people respond differently to synthesised voices depending on how they sound and whether they are polite [12, 13].

While voice interaction is associated with numerous benefits, literature also looks at several negative aspects. Researchers have investigated different privacy concerns of using digital voice assistants [42]. This research has led to studies that aim to mitigate potential attacks, such as the work by Kwak *et al.* [41] that distinguish genuine voice commands from potential voice based attacks. In addition, voice interaction is not considered socially acceptable in all public situations [53].

Crowdsourcing with Audio and Speech Data

There exists a wide range of crowdsourcing tasks that use speech or audio data [21]. Such tasks require workers to listen to audio data and/or provide answers through voice input. For instance, crowdsourcing has been used to gather speech data from different local dialects [43], rate speech data for assessing speech disorders [7], annotate audio data [22, 25], and annotate speech data for training automatic speech recognition systems [8]. In a speech sound rating task, Byun *et al.* [7] state that the inability to standardise equipment or playback is a major limitation when using an online crowdsourcing platform like Amazon Mechanical Turk. There are also numerous other tasks, such as sentiment analysis and moderation, that can be completed via voice input although they typically contain text data and text responses.

Vashistha *et al.* [61] introduced 'Respeak', a mobile application that uses voice input for crowdsourcing speech transcription tasks. In the study, participants listen to short audio clips and repeat what they had heard. In a deployment with 25 university students in India, the study shows that audio files could be transcribed with a word error rate of 8.6% for Hindi and 15.2% for Indian English. The application uses Google's Android Speech Recognition API to generate transcripts of user utterances. An extension of the proposed application was also successfully used to crowdsource speech transcription tasks from visually impaired users [62] and through basic phones [60]. However, all three studies are limited to speech transcription and none of them are fully functional hands-free voice interfaces that have the capability to browse available tasks, launch tasks, and check progress.

In a vision paper, Hettiachchi *et al.* [33] propose that it is feasible to use smart speakers for crowdsourcing and discuss potential benefits like low cost of entry, ubiquitous nature,

efficiency and accessibility. They also highlight several challenges such as privacy concerns, integration issues, and impact on data quality when multitasking. We extend this work with an empirical evaluation, where we present a functional voice interaction application for crowdsourcing with several different tasks, and evaluate the system using both a lab study and a field deployment.

Crowd Worker Context

In most crowdsourcing platforms, such as MTurk, Figure Eight, and Prolific, crowd workers actively select and launch tasks they wish to work on. This model typically introduces higher latencies for tasks that require workers with specific skills (*e.g.,*Translation) [20]. As a solution, several studies have investigated the possibility of proactively delivering tasks to workers instead of waiting for them to initiate the task [1, 36]. In mobile crowdsourcing, Acer *et al.* [1] investigate how worker mobility patterns, workflow, and behavioural attributes can be used to identify opportune moments to deliver tasks to mobile crowdworkers. The study aims to embed crowdsourcing tasks to workers' daily routine and reports increased worker response rate and accuracy. In crowdsourcing, task requesters also aim to capture the cognitive surplus of workers, which is described as the free time of individuals who are capable of contributing to a task [55]. Different techniques can be used to tap into the cognitive surplus. Goncalves *et al.* [27, 29] show that interactive public displays can be successfully used to gather input from people who are idling at public spaces, while Hosio *et al.* [35] demonstrated the feasibility of a situated crowdsourcing system. In another example, Skorupska *et al.* [57] show that older adults can contribute to a transcription task while watching a movie.

By using digital voice assistants for a broader spectrum of crowd tasks, we aim to reduce the complexity of initiating crowd work. By doing so, this is likely to lead to a better utilisation of cognitive surplus and opportune moments for crowdsourcing purposes.

CROWD TASKER SYSTEM

To enable crowdsourcing through digital voice assistants, we developed Crowd Tasker, an application for Google Assistant which prompts crowd tasks to users and stores responses. We opted for Google Assistant as it has the largest market share in Digital Voice Assistants [50], and allows us to easily deploy our application to both smart speakers and smartphones. We used Dialogflow[2] and the NodeJS client library for Actions on Google[3] to process user utterances and manage the crowd task flow. Using Dialogflow we mapped users' voice input to a set of pre-configured intents that lead to different actions. An intent represents an end-user's intention for one conversation turn. It also allowed us to activate different intents based on the context, such as a previous response by the user. Figure 1 shows the different intents we developed considering main use cases of online crowd work along with their flow within the application. We iteratively improved our system prototype to provide a unified user experience.

[2]https://dialogflow.com/

[3]https://github.com/actions-on-google/actions-on-google-nodejs

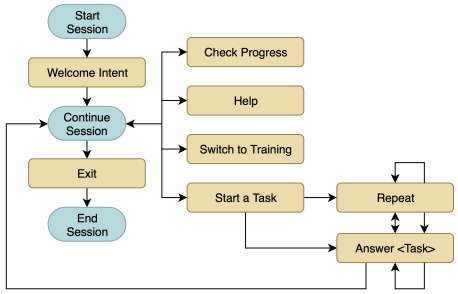

Figure 1: Intents and session flow of Crowd Tasker system.

Our application works with Google Assistant, and users can launch it by saying 'Talk to Crowd Tasker'. The application does not use any visual prompts and can be initiated from any device that supports Google Assistant (*e.g.,* smartphone, smart speaker, smart watch). Users can check what kind of tasks are available, start a particular task, review their progress, and ask to repeat questions. The system also asks users to provide a clear response or seek help and listen to instructions when it cannot understand the user's utterance. Below is an example for a conversation between Crowd Tasker and a user.

Participant : Okay Google, talk to Crowd Tasker
Crowd Tasker: Hi, I am the Crowd Tasker, What can I do for you today?
Participant: What kind of tasks do you have?
Crowd Tasker: There are six available tasks: Sentiment analysis, Comprehension, [...] and Emotion labelling. Which task do you want to complete?
Participant: Please start Audio Annotation task.
Crowd Tasker: <audio clip playing>
Participant: A dog barking
[Conversation continues]

Crowdsourcing Tasks

We created six unique crowdsourcing tasks under two categories: voice-compatible (*i.e.,* typically text-based tasks, but can be adapted for voice input) and voice-based (*i.e.,* audio or speech related). The voice-compatible tasks are sentiment analysis, comprehension, and text moderation. The voice-based tasks are audio annotation, speech transcription and emotion labelling. Task selection was informed by a crowd task taxonomy [24] and task availability [19] from prior work to be representative of typical tasks available in crowdsourcing platforms. Each task contained 10 different questions with varying complexity extracted from datasets that are extensively used in the literature.

Sentiment Analysis: Participants are asked to identify the sentiment of a given sentence. We adopt the task from prior work that investigates crowd task performance [26, 28, 34]. Sentiment can be classified as either 'positive', 'negative' or 'neutral'.

Comprehension: Participants provide an answer to a question based on a provided text. Questions are based on different Wikipedia articles. For the task deployed through the voice assistant, we use English speech data from The Spoken Wikipedia project [40].

Text Moderation: Workers are asked to label text messages as 'spam' or 'not spam'. Data is extracted from the SMS Spam Collection [3]. In the web interface, the message is presented in text format, whereas when using the voice assistant, participants listen to the message as generated by Google's text-to-speech service.

Audio Annotation: In this task, participants are asked to provide a label that describes a sound they hear. All the audio clips and ground-truth labels are extracted from the Freesound Data set [22]. An answer is considered accurate if it matches any of the valid keywords for the clip. For example, for a clip of a moving horse carriage, terms such as horse and cart are considered as valid answers.

Speech Transcription: In the speech transcription task, participants listen to a short audio clip (average length of 3 seconds) which contains an utterance of an English speaker. Participants are asked to clearly speak out or type in what they heard. Speech data and transcripts are sourced from the Noisy speech database [59]. We use the Levenstein distance [17] to calculate the accuracy of each answer.

Emotion Labelling: For the emotion labelling task, we use the Multimodal EmotionLines Dataset [51], which contains short utterances of different people from a popular TV show. We extract audio clips and ground-truth labels for two people. Workers are asked to categorise the emotion of each utterance as either 'anger', 'disgust', 'fear', 'joy', 'sad', or 'surprise'.

STUDY

Lab Study

We conducted a lab study to compare crowd task performance through web (*i.e.,* using a regular graphical user interface) vs. digital voice assistants (*i.e.,* using a voice interface). Hence, we also built a simple web application that replicates the task completion interface of a typical crowdsourcing platform. The system was developed using Python (Django framework) and connected to the Crowd Tasker database that contains task and performance data.

We recruited 30 participants through a university-wide online notice board, using two eligibility constraints: we only recruited native or fluent English speakers, and only participants who have used digital voice assistants. During screening, participants reported whether they used digital assistants frequently (daily or more than few times a week), occasionally (few times a week), or rarely (few times a month). We balanced the use of voice assistants by recruiting 10 participants for each category (30 participants in total). Participants were compensated with a $20 gift voucher.

Participants completed tasks under two conditions: using a desktop-based web interface (Figure 2), and a Google Home smart speaker. Initially, participants completed a training round in which they completed one question from each task for both conditions. We then counter-balanced the order of the experimental conditions, and randomised the completion order of all tasks a priori. Each task contains 10 questions, and participants answered 5 questions per condition. Table 1 summarises tasks under each condition. Finally, participants completed a short exit interview to discuss their experience, and we probed them about convenience, perception of the two conditions, and task difficulty.

	Question		**Answer**	
Task	*Web*	*VA*	*Web*	*VA*
Sentiment Analysis	Read	Listen	Button	Speak
Comprehension	Read	Listen	Type	Speak
Text Moderation	Read	Listen	Button	Speak
Audio Annotation	Listen	Listen	Type	Speak
Speech Transcription	Listen	Listen	Type	Speak
Emotion Labelling	Listen	Listen	Button	Speak

Table 1: Crowdsourcing Tasks

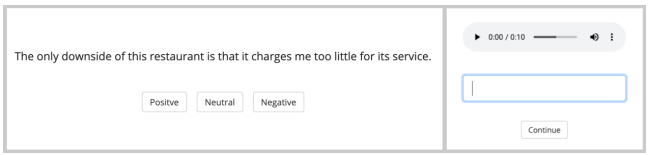

Figure 2: Screenshots of the web interface for Emotion Labelling (left) and Comprehension (right) tasks

Field Deployment

To further examine the feasibility of using digital voice assistants for crowdsourcing purposes we conducted a field deployment. Before proceeding with the field deployment, we made several enhancements to the system based on the feedback of participants of the lab study, including several workflow improvements. For instance, in the lab study, participants listened to the text segment prior to the question. We swapped the presentation order, so that participants could anticipate the relevant information when listening to the text. We also improved the timing for gaps in between spoken text to enhance the quality of overall conversation experience. For example, in the lab study, participants mentioned that they found it difficult to distinguish between instructions and the first question of a task due to absence of an appropriate time gap (similar to proximity in Gestalt Principles [30]. Finally, we added an intent, that allows participants to check their progress and know how many questions are remaining in each task.

We recruited 12 participants through our university's online notice board. Similar to the lab study, we set out eligibility constrains and recruited a population that is balanced in terms of participants' experience with voice assistants. Additionally, we did not recruit any of the participants who completed the lab study.

At the beginning of the study, we met our participants in person, provided them a Google Home Smart Speaker, and asked them to use it in their home for a period of 7 days. We also instructed participants on how to setup the device and use the application, and finally gave them a brief demonstration. We asked them to complete all the training tasks first. We then explained how they could complete tasks: through the provided smart speaker or using the digital voice assistant application on another device.

In the field deployment, participants were required to complete 6 tasks similar to the lab study. To replicate the reward mechanism of a standard crowdsourcing marketplace, we informed the participants that they would be compensated based on the number of tasks they complete in the study. Participants were given a gift voucher of up to $30 if they completed all available tasks. After a week, participants returned the smart speaker and took part in a short interview about their experience. We asked participants to report on the level of convenience, whether they were doing any other activity while completing tasks, and how they compare interacting through the smart speaker and another device.

RESULTS

Lab Study

30 participants (18 women and 12 men) completed the lab study. Participant age ranged from 18 to 38 years ($M = 25.7$, $SD = 5.7$). 8 participants were native English speakers, while the remaining participants were fluent English speakers. We did not observe any significant impact on task accuracy in terms of participant demographics (age and gender) or voice assistant usage. We also asked participants whether they had prior experience with particular voice assistant services such as Google Assistant, Siri, and Alexa. However, there was no significant effect on task performance from any of the indicators.

Web Interface vs. Voice Assistant

Differences in worker performance between the web interface and the voice assistant in terms of accuracy and task completion time are shown in Figure 3. A paired-sample t-test indicates that there is no significant difference in accuracy between the web and voice assistant conditions for the text moderation task ($t(29) = -0.90$), $p = 0.38$). Similarly, a Wilcoxon signed rank test revealed that there is no difference in task accuracy for the emotion labelling task ($Z = 89$, $p = 0.90$). Task accuracy is significantly higher in the web interface for all the remaining tasks.

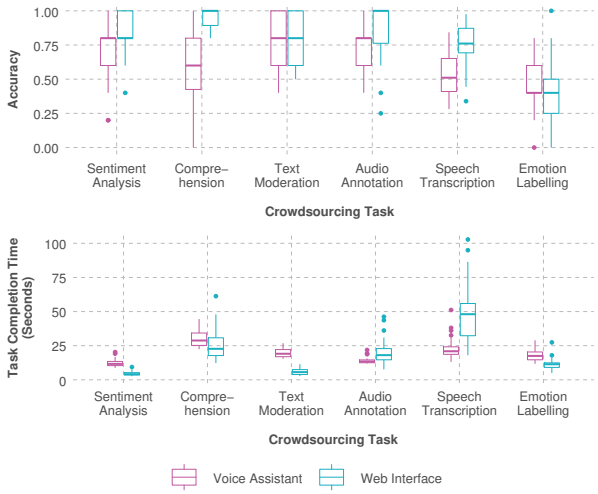

Figure 3: Worker accuracy (top) and task time (bottom) across crowd tasks when using the web and voice assistant

Paired t-tests indicated that task completion time is significantly lower in the voice assistant condition for two voice-based tasks, audio annotation ($t(29) = -3.62$, $p < 0.01$) and speech transcription ($t(29) = -6.33$, $p < 0.01$). For all 3 voice-compatible tasks and emotion labelling task, task completion time in voice assistant is significantly higher than the web interface.

Native English Speakers

As shown in Figure 4, when completing tasks through voice, native English speakers exhibit a higher task accuracy for most tasks when compared to the remaining participants.

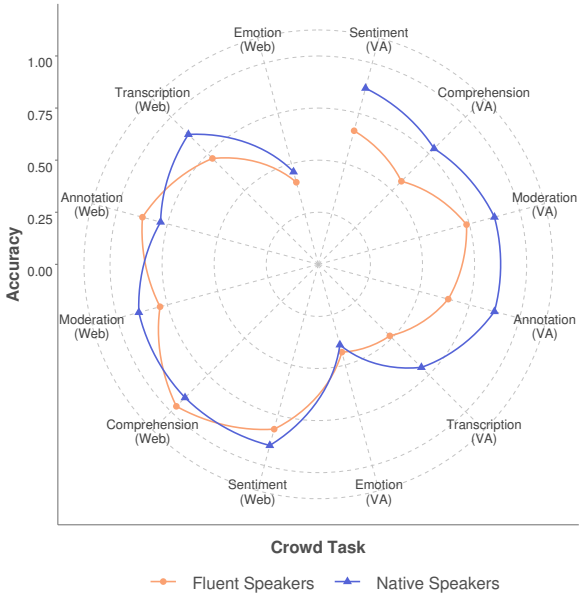

Figure 4: Native and fluent English speakers accuracy across crowd tasks when using voice assistants and web

We further examined the difference in task accuracy for native speakers. A paired-sample t-test was conducted to compare the accuracy in web and in voice assistant conditions. There was no significant difference in the accuracy for web and voice assistant conditions in sentiment analysis task ($t(7) = -0.42$, $p = 0.68$), moderation task ($t(7) = -0.11$, $p = 0.91$), audio annotation task ($t(7) = 0.77$, $p = 0.46$), and emotion labelling task ($t(7) = -0.39$, $p = 0.71$). As the accuracy scores of reading comprehension tasks were not normally distributed, we used a Wilcoxon signed rank test and found no significant difference in accuracy in web and voice assistant conditions ($Z = 1$, $p = 0.42$).

Native English speakers exhibited a similar variation to the general sample, considering task completion time. To compare the task completion time in the web interface and voice assistant, we conducted paired t-tests or Wilcoxon signed rank tests (when the sample was not normally distributed). For speech transcription, task completion time was significantly lower in voice assistant condition ($t(7) = -3.42$, $p = 0.01$). For audio annotation, there was no significant difference between two conditions ($t(7) = -1.05$, $p = 0.33$). For emotion labelling task and the remaining voice-compatible tasks, task

completion time was significantly higher in voice assistants when compared to web interface. Results from our lab study, for the general sample and the native English speakers are summarised in Table 2.

	Accuracy		Task Time	
Task	*General Sample*	*Native Speakers*	*General Sample*	*Native Speakers*
Sentiment Analysis	↓	-	↑	↑
Comprehension	↓	-	↑	↑
Text Moderation	-	-	↑	↑
Audio Annotation	↓	-	↓	-
Speech Transcription	↓	↓	↓	↓
Emotion Labelling	-	-	↑	↑

Arrows indicate that the measure is significantly higher (↑) or lower (↓) in the voice assistant when compared to the web interface. No statistically significant difference between conditions is indicated through a dash (-)

Table 2: Summary of statistical results of the lab study

Qualitative Data

To further extend these results, we present a qualitative analysis of our semi-structured interview data. Informing ourselves through the aforementioned quantitative results and the setup of the field study, we apply the general inductive approach to data analysis as defined by Thomas [58]. Two of the paper's authors (one of which conducted the interviews) independently analysed and coded the interview data – after which three of the authors agreed on the final set of themes. Given the study's exploratory character, we focus on the following themes; 'participant interaction', 'task suitability', and 'perceived usefulness'.

Participant interaction

Participants considered the specific advantages of using either a web interface or a voice assistant. The web interface was perceived as offering participants a higher level of control over their input, with one participant describing this as;

> P08: "I feel I am more accurate and precise when I type. I also feel I'm more in control in the web."

The web interface allowed participants to work at their own pace, without pressure from a timeout by the voice assistant.

> P10: "Voice has more pressure, I need to give a timely response. [Using the web interface,] I feel more relaxed as the pace is defined by me."

However, a number of participants ($n = 13$) found it easier, more efficient, or simply more enjoyable to speak out the answer as opposed to typing. Participants compared the interaction with the voice assistant to be more human-like as compared to input provided through a web interface.

> P10: "Also when the answer is too long then voice is easier because it saves time of typing."

Furthermore, the voice interface provided participants with benefits we did not initially consider. One participant stated that the use of voice commands requires less focus on the correct spelling of words.

> P15: "If you want me to type, I am not sure about the words (spellings), but I don't have to worry about that when speaking."

Task suitability

Given the wide range of tasks included in our study, we aimed to identify task suitability in relation to voice interaction. Several participants ($n = 10$) reported that they found it difficult to remember content when completing tasks through the voice assistant. Participants also mentioned that they were unable to memorise all options in the emotion labelling task.

> P18: "Speech transcription and comprehension tasks were harder on voice assistant, because I had to remember longer sentences."

> P16: "Emotion labelling was difficult because I didn't remember the emotions."

Therefore, when interacting through voice, participants preferred tasks with fewer options such as text moderation and sentiment analysis and tasks with short answers like audio annotation. A number of participants also mentioned that the use of voice control allowed them to respond quicker and accelerate the interaction.

> P24: "The rest of the tasks (apart from comprehension) were easier with the speaker, because it took off the effort of reading and typing manually."

Perceived usefulness

Although the study was carried out in a controlled lab environment, we were interested in the participants' perceived usefulness of a voice assistant in completing crowdsourcing tasks. Participants believed that the use of a voice assistant would enable them to simultaneously work on something else.

> P21: "Using the voice assistant, you can be busy and multitask but with web its not possible."

Furthermore, participants described scenarios in which they believe the use of a voice assistant would be useful, including examples such as leisure time, cleaning, and cooking.

> P26: "I think I will use voice assistant a lot during my leisure time. During the weekend I do household work so I can use the speaker."

Field Deployment

12 participants (5 women and 7 men) aged between 20 to 32 years old ($M = 26.7$, $SD = 3.8$) completed the field deployment study. Five participants were native English speakers, while the remaining participants were fluent English speakers. All participants completed the full set of tasks within the one week period.

In Figure 5, we can observe a similar task performance in the lab study and the field deployment. Our statistical analysis confirmed that for all the tasks, there is no significant difference in both the accuracy and task time between tasks completed through the smart speaker in the lab study and the field deployment.

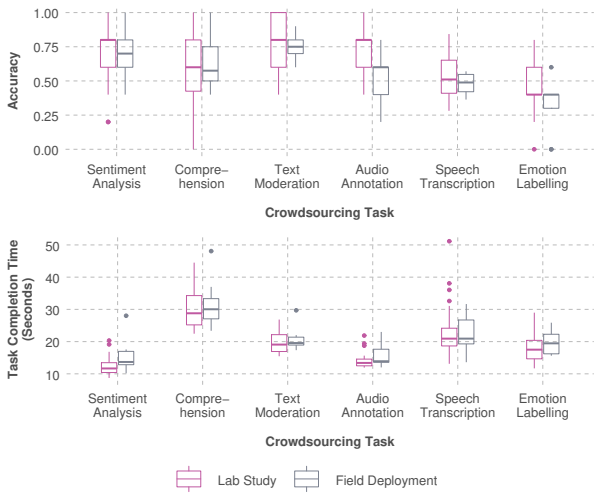

Figure 5: Task accuracy (top) and Task completion time (bottom) for participants in the lab study and in field deployment

To understand participants' voice assistant usage for crowd work, we further examined their usage patterns from the task completion data. Figure 6 shows the total number of question answered by all participants over the time of day.

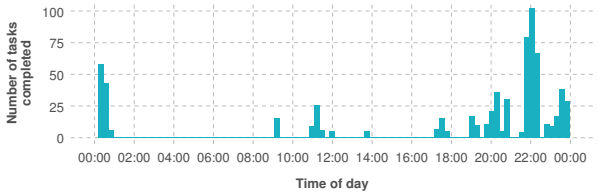

Figure 6: Total number of questions answered by participants over the time of the day

As exemplified in Figure 7, six of the participants completed tasks over more than one day. The other six participants completed all questions within a single day. Figure 8 shows the question completion over time for those participants. Although they completed all questions within the same day, we notice that they used multiple brief sessions to complete tasks with interruptions or breaks in between.

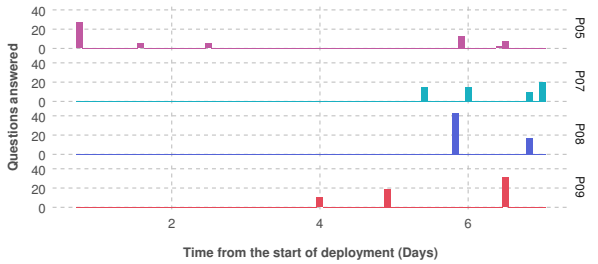

Figure 7: Task completion by day of deployment for four participants

While all participants had a smart speaker set up in their home, they were also given the option to complete tasks through digital voice assistants on their smartphones. In the field

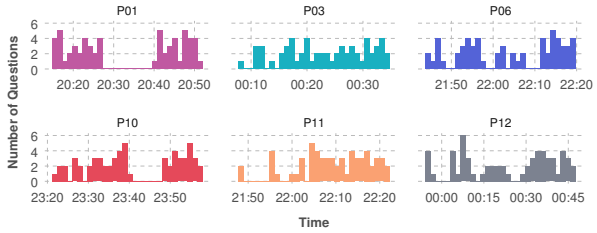

Figure 8: Task completion by time of day for six participants

deployment, 7 participants completed all tasks through smart speakers, whereas 3 participants used the smartphone for all the tasks. Only 2 participants used both devices, completing 35 and 30 questions through their smartphone.

Application logs indicate that participants used 129 sessions in total to interact with Crowd Tasker. On average participants used 13.07 queries per session. Table 3 presents a summary of user utterances indicated by corresponding intents matched by voice assistant during the field deployment. We notice a high number of repeats for Speech Transcription task. 'Check progress' intent has the highest number of matches after 'Start a task intent' and a higher exit rate, suggesting participants often checked their progress before closing the application.

Intent	*Num. of sessions*	*Num. of matches*	*Exit percentage*
Welcome Intent	129	130	13.1%
Switch to Training	24	31	9.7%
Start a Task	80	241	4.6%
Check Progress	59	142	16.2%
Check Available Tasks	21	31	6.5%
Sign-in (during briefing)	12	12	100%
Help	3	3	0%
Answer - Sentiment Analysis	26	129	3.9%
Answer - Comprehension	23	126	3.2%
Answer - Text Moderation	17	120	0%
Answer - Audio Annotation	22	128	0%
Answer - Speech Transcription	20	122	1.6%
Answer - Emotion Labelling	39	135	3.7%
Repeat	6	17	5.88%
Repeat (after starting a task)	21	43	9.3%
Repeat - Speech Transcription	13	55	1.8%
Repeat - Comprehension	6	24	0%
Repeat - Audio Annotation	2	3	0%

Table 3: Summary of intent matching

Qualitative Data

Similar to the qualitative analysis of our lab study, we again perform an inductive approach to the analysis of our semi-structured interviews obtained following the field deployment [58]. The focus of our analysis is on the practical aspects of crowdsourcing using a voice assistant *in situ*. We therefore focus on the following themes; 'ease of use', 'multitask behaviour', and the use of 'smartphone vs smart speaker'.

Ease of use

Our in-the-wild deployment highlighted issues in interacting with the device. For example, the voice assistant could occasionally not recognise participants' utterances due to accent, background noise, or volume level. An issue we previously did not consider was the interaction between participant devices. Some participants reported that multiple devices (*e.g.,*both smartphone and smart speaker) were activated when issuing the command to initiate the voice assistant.

> P07: "My phone got activated when I spoke to the speaker."

Furthermore, the level of voice recognition differed between participants, occasionally hampering the participant's ability to complete tasks.

> P03: "Sometimes the speaker had trouble understanding my accent or it didn't pick up my voice."

Generally, however, our participants ($n = 8$) highlighted that completing tasks through voice was convenient and the system was easy to use. In particular, the launching of tasks was seen as straightforward in comparison to the use of a computer.

> P02: "It was quick and easy to complete tasks through the voice assistant."
>
> P10: "Good thing about voice assistant is that I could quickly start tasks."

Multitask behaviour

Participants mentioned that they had to change their attention depending on the task. These participants ($n = 7$) were able to attend to other tasks or switch context while going through tasks.

> P09: "I was interrupted once. I had to talk to my flatmate."
>
> P06: "I was folding laundry while doing the tasks. I really didn't have any problem."

Some participants ($n = 3$) also mentioned that they initiated Crowd Tasker at opportune moments such as during idling, followed by a routine task, or when they needed a distraction from a particular task.

> P03: "I was free when I started it (Crowd Tasker) and I was sitting on a couch. I was occasionally checking my smartphone for notifications while doing the tasks."
>
> P02: "Probably I did after dinner. I was paying full attention but watching something in between tasks."
>
> P07: "Even when I was using the speaker, I was helping my friend arrange the house. I was also in the middle of an assignment. Because I wanted a distraction, I started the task."

Smartphone vs Smart Speaker

Participants who opted to use their smartphone instead of the smart speaker appreciated the fact that they can get a visual confirmation of their utterances.

> P06: "If I am using the phone, I know if Google understood me correctly."

Participants also highlighted that it was beneficial to see the question on the screen and then answer through voice when the task was too complex.

> P11: "Half of these tasks are easy with voice. For others it might be good to use voice assistant in phone, so you can see the question but still can answer through voice."

As most participants ($n = 7$) chose to use the smart speaker for all the tasks, they commented on positives of the speaker such as better audio quality and voice recognition distance as compared to smartphones.

DISCUSSION

Our study is the first to systematically investigate the possibility of using voice-only interfaces, such as smart assistants, for crowdsourcing purposes with a variety of tasks. Through a lab study and a field deployment we are able to demonstrate the feasibility of this approach, and at the same time highlight a number of remaining research and design challenges.

Our work is at the nexus of the literature on crowdsourcing and voice interaction. These have been largely distinct, each having a long tradition of design guidelines, best practice suggestions, and research findings. As we discuss here, we find that the usability of voice interaction needs to be carefully thought through when developing voice interfaces for crowd work, particularly by taking into account the nature of crowd work. For instance, we find that crowd work requires humans to provide answers to the agent's questions, whereas typically with voice assistants it is humans who provide the questions and agents the answers. This poses a number of challenges that we highlight in our discussion.

Crowdsourcing through Voice

We show that crowd tasks can be completed through a voice assistant with an acceptable level of speed and accuracy, and that a voice assistant can provide crowd workers with greater flexibility in how they approach crowd tasks compared to a regular web interface. Indeed, participants were faster at completing free-form answer tasks with Crowd Tasker than with a typical web interface. Prior work reports that high quality data could be obtained by using voice input for crowdsourcing speech transcription tasks [60, 61, 62]. While we were able to obtain reasonable data quality for our transcription task, as detailed in Table 2, results from our lab study indicate that other tasks are better-suited for voice-based crowdsourcing systems. For crowdsourcing platforms, it would be more productive to create a unified system that issues online and voice-based crowd work depending on the type of task. Perhaps, Amazon Web Services is best positioned to achieve this as they own a crowdsourcing platform (Mechanical Turk) as well as a voice assistant service (Alexa). Other crowdsourcing platforms can also plausibly create voice assistant applications that tie into their own market. This would also require further research that explores dynamic task assignment to either a web or voice interfaces.

In contrast to previous systems that involve calling a phone number to access an interactive voice response (IVR) application [60] or providing voice input through a smartphone appli-

cation [61, 62], accessing CrowdTasker through an always-on microphone can bring numerous benefits to users. The qualitative results of the field deployment highlight that participants were able to utilise their cognitive surplus by initiating tasks at opportune moments [55]. The hands-free interaction through the speaker also allowed participants to multitask while completing crowd tasks. In Figure 6, we observe that participants completed most tasks outside regular working hours. We also note that there is no statistically significant difference in performance with the voice assistant between the lab study and the field deployment, suggesting that multitasking did not have a negative impact on the data quality.

Although we recruited fluent English speakers for our study, we observed a significant difference in task accuracy between native and non-native English speakers. Our findings are in line with prior work that states that native English speakers have a better overall experience with smart speakers [52]. From our qualitative analysis, we understand that this difference is due to recognition problems on both sides of the interaction: the voice assistant being less successful at recognising non-native English speech, and the non-native English speakers being less able to comprehend the voice assistant's speech due to tone, accent, and speech rate. Therefore, when using voice assistants for crowd work, language competency of the worker will play an important role. Using language-based pre-selection mechanisms or support for multiple languages could be feasible solutions to mitigate this factor. We also note that language skills are important for a wide array of crowdsourcing tasks in regular web-based crowdsourcing platforms, so this problem is not unique to the voice interface.

Developing Voice-based Crowdsourcing Platforms

While our study reveals promising results for the use of voice assistants for crowd tasks, it also identifies several factors that undermine the user experience and data quality. We discuss these challenges and propose ways to address them in the development of future voice-based crowdsourcing systems.

Optimising workflow to provide control for workers

Our participants mentioned that they felt less in control when using the voice assistant. This is consistent with prior studies that have found voice interaction to be associated with a lower subjective sense of control in both smart-home interfaces [45] and digital games [2]. We propose three features to reduce this perceived lack of control. First, the voice assistant should repeat the entire or part of a task question upon request. Crowd Tasker gives the user an option to repeat an entire question, but the qualitative feedback from our study highlights that this feature should be extended to provide more granular control. For instance, each question in the comprehension task consisted of two parts: a question and a sentence. In the current implementation Crowd Tasker repeats both parts when asked, forcing the user to listen to seconds of potentially irrelevant content when they may only want to check a single word. In a future implementation, it would be useful for the user to be able to request only a specific part to be repeated.

Second, workers should be able to stop and resume tasks at any point. Although Crowd Tasker was designed to provide a predefined number of questions in each task, it will exit the application if it receives no response from the worker after repeated prompts. When the worker returns to the voice assistant, they can resume from the last completed question. For voice-based crowd work, we recommend including more task checkpoints than in web-interfaces.

Third, for certain task types, it is useful if the worker can skip certain sections or interrupt the voice assistant while it is speaking. For example, in the text moderation task, workers had to listen to the entire message even if they had figured out their answer within the first few words. Currently, popular commercial voice systems provide limited interruptibility, but future systems that allow for more interruptions are likely to provide a more pleasing user experience for crowdsourced voice work.

Finally, due to the nature of voice interaction, it can be beneficial to ask participants a question first, and then provide them the relevant stimuli to complete the task. In screen-based crowd work, this information is often given in the reverse order, but due to the visual nature of those tasks it is possible for workers to quickly switch between task description and stimulus. With a voice-only modality, our participants preferred to be asked the question first, so that they know what to look for when listening to a stimulus, effectively having a reduced working memory load. This ultimately reduces the need to repeat the task description, improve accuracy, and result to greater subjective satisfaction.

Handling responses

When designing and developing Crowd Tasker, handling responses for questions that require free-form answers proved to be particularly challenging. We also note that during the field deployment, specific worker commands such as asking to repeat the question were, on occasion, erroneously captured as answers. As a possible solution to mitigate errors, we propose that future systems should allow users to listen to and revert their answer if necessary.

Payment

The payment mechanism in a voice-based crowdsourcing system can be similar to existing online crowdsourcing platforms where workers are paid per completed task [18], with each task having a maximum time limit. In a voice-based system, prompts should be added to indicate if a worker runs out of time. When estimating task times for payments, it is important to consider the time taken to playback the question or prompt, to ensure fair compensation [31].

Task allocation and recommendation

In conventional crowdsourcing platforms, workers need to browse and select the task they wish to attempt. This process takes a considerable amount of time and effort [11]. Similarly, browsing through a large number of tasks is not desirable in a voice interface. Therefore, it is critical to allocate or recommend a handful of relevant tasks to workers when delivering tasks through voice. There is a large body of work that analyses different worker attributes [34, 37], and behavioural traces [23, 54] that can be utilised to match tasks to workers. In addition, our participants mentioned that they had to repeatedly ask for available tasks as they had to use the specific task

name to initiate the task. Thus, a feature that automatically starts a relevant task for the user will be particularly useful in a voice-based crowdsourcing system.

Selecting tasks for voice

Our findings show that certain tasks are more appropriate for voice interaction than others. We initially anticipated that inherently voice-based tasks (such as audio annotation, speech transcription, and emotion labelling) would be more suitable than tasks that are voice-compatible but text-based (such as sentiment analysis, comprehension and text moderation). However, our analysis suggests that other factors play a more critical role in determining task suitability, such as demand on working memory and task complexity. We discuss such factors extensively in the following section.

Designing Voice Interaction for Crowd Tasks

While there exists a significant amount of research regarding conversational interfaces, our study shows that crowd work is a peculiar case. Whereas in traditional conversational interaction the user may be prompted to talk about their desires and preferences, in the crowd work scenario users are typically prompted to talk about stimuli they have just heard, which increases their cognitive load. Minimising strain on users is crucial, as satisfaction with voice assistants has been shown to vary according to the level of effort involved in the tasks for which they are used [38]. Therefore, based on our results, we highlight several important considerations for designing voice systems for crowd work.

Shortening questions and answers

Voice interfaces often place higher demand on users' short-term and working memory compared to graphical interfaces, due to both the lack of visible cues and the fact that speech uses more of these cognitive resources than hand-eye coordination does [56]. This was manifested in our study, as participants struggled with those tasks that required them to hold information in mind while completing an action through voice. Accuracy was decreased and subjective reported effort was increased for Speech Transcription and Comprehension in particular, as these tasks involved working with samples of speech that were too long to hold in working memory. Hence, we suggest as a general rule that the amount of information presented in a single conversation turn of a voice-based crowd task should be kept to a minimum.

The same holds for longer answers. For tasks such as Speech Transcription, which required long responses, participants preferred to type out the recording using the web interface rather than dictate it by voice, as typing made it easier to transcribe short chunks at a time, rather than attempting to transcribe the entire sentence at once. Therefore, in addition to short prompts and questions, it is also important to keep the user's required responses as short as possible when designing crowd tasks for voice. When it is necessary to work with longer sentences, such as in transcription, it is recommended to break them into smaller sub-tasks [10].

Reducing the number of options in multi-label tasks

Our study had three tasks that required participants to choose a predefined option as the answer. Of these, participants found the emotion labelling task particularly difficult as the voice condition required them to remember 6 different response options. In contrast, many participants mentioned the sentiment analysis (options: 'positive', 'negative', 'neutral') and text moderation (options: 'spam', 'not spam') tasks as being no more difficult with voice than with the web interface. Therefore, we recommend that voice-based crowd tasks should provide only a small set of options for the user to choose from. Where multiple labels are necessary, a task decomposition technique could be used to transform the multi-label task into multiple binary labelling tasks [63]. For some tasks, such as emotion labelling, it may be feasible for the system to obtain an open answer, and then map that answer on to a hidden set of options using natural language processing techniques.

Limitations

We acknowledge three limitations in our study. First, our evaluation is limited to one voice assistant. While there are several other services available, we decided to use Google Assistant due to the complexity introduced to the study design when using multiple services. In addition, Google Assistant currently holds the largest market share [50] and there is currently no substantial difference in workflow or the ability to recognise voice across the major services.

Second, our participants are not regular crowd workers and have only limited experience with crowdsourcing. While it would be more relevant to evaluate this system with crowd workers, there are numerous practical difficulties in deploying a system of this nature in the wild.

Third, our field deployment is limited to a week as we used the same questions from the lab study for better comparability. A more longitudinal future study that recruits more participants can reveal further insights on designing commercial voice-based crowdsourcing systems.

CONCLUSION

To investigate the feasibility of using digital voice assistants for crowdsourcing, we developed Crowd Tasker, a novel stand-alone crowdsourcing system that runs on Google Assistant. Through a lab study, we report that for native English speakers, there is no significant difference in task accuracy when completing five types of tasks through a regular web interface and a voice interface while task completion time varies depending on the task. Further, through a field deployment, we show that participants were able to complete tasks conveniently through voice at their home. They were able to multi-task, launch tasks at opportune moments, and resume their work if interrupted or distracted while using the voice assistant.

We identify several approaches on how to optimise voice-driven workflow, handle responses, recommend or allocate tasks, and select tasks when developing voice-based crowdsourcing systems. We anticipate that our work will lay the foundation for different research avenues to explore voice-based crowd work as a noteworthy addition to the existing crowdsourcing eco-system, and help create more accessible and convenient platforms for crowd workers.

REFERENCES

[1] Utku Günay Acer, Marc van den Broeck, Claudio Forlivesi, Florian Heller, and Fahim Kawsar. 2019. Scaling Crowdsourcing with Mobile Workforce: A Case Study with Belgian Postal Service. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 2, Article 35 (June 2019), 32 pages. DOI: http://dx.doi.org/10.1145/3328906

[2] Fraser Allison, Joshua Newn, Wally Smith, Marcus Carter, and Martin Gibbs. 2019. Frame Analysis of Voice Interaction Gameplay. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 393, 14 pages. DOI: http://dx.doi.org/10.1145/3290605.3300623

[3] Tiago A. Almeida, José María Gómez Hidalgo, and Tiago P. Silva. 2013. Towards SMS Spam Filtering: Results under a New Dataset. *International Journal of Information Security Science* 2, 1 (2013), 1 – 18.

[4] Frank Bentley, Chris Luvogt, Max Silverman, Rushani Wirasinghe, Brooke White, and Danielle Lottridge. 2018. Understanding the Long-Term Use of Smart Speaker Assistants. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3, Article 91 (Sept. 2018), 24 pages. DOI:http://dx.doi.org/10.1145/3264901

[5] Janine Berg. 2015. Income security in the on-demand economy: Findings and policy lessons from a survey of crowdworkers. *Comp. Lab. L. & Pol'y J.* 37 (2015), 543.

[6] Michael Braun, Anja Mainz, Ronee Chadowitz, Bastian Pfleging, and Florian Alt. 2019. At Your Service: Designing Voice Assistant Personalities to Improve Automotive User Interfaces. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 40, 11 pages. DOI: http://dx.doi.org/10.1145/3290605.3300270

[7] Tara McAllister Byun, Peter F. Halpin, and Daniel Szeredi. 2015. Online crowdsourcing for efficient rating of speech: A validation study. *Journal of Communication Disorders* 53 (2015), 70 – 83. DOI: http://dx.doi.org/10.1016/j.jcomdis.2014.11.003

[8] Chris Callison-Burch and Mark Dredze. 2010. Creating Speech and Language Data with Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (CSLDAMT '10)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1–12.

[9] Veena Chattaraman, Wi-Suk Kwon, Juan E. Gilbert, and Kassandra Ross. 2019. Should AI-Based, conversational digital assistants employ social- or task-oriented interaction style? A task-competency and reciprocity perspective for older adults. *Computers in Human Behavior* 90 (2019), 315 – 330. DOI: http://dx.doi.org/10.1016/j.chb.2018.08.048

[10] Justin Cheng, Jaime Teevan, Shamsi T. Iqbal, and Michael S. Bernstein. 2015. Break It Down: A Comparison of Macro- and Microtasks. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 4061–4064. DOI: http://dx.doi.org/10.1145/2702123.2702146

[11] Lydia B. Chilton, John J. Horton, Robert C. Miller, and Shiri Azenkot. 2010. Task Search in a Human Computation Market. In *Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP '10)*. ACM, New York, NY, USA, 1–9. DOI: http://dx.doi.org/10.1145/1837885.1837889

[12] Leigh Clark. 2018. Social Boundaries of Appropriate Speech in HCI: A Politeness Perspective. In *Proceedings of the 32nd International BCS Human Computer Interaction Conference (HCI '18)*. BCS Learning & Development Ltd., Swindon, UK, Article 76, 5 pages. DOI:http://dx.doi.org/10.14236/ewic/HCI2018.76

[13] Leigh Clark, João Cabral, and Benjamin Cowan. 2018. The CogSIS Project: Examining the Cognitive Effects of Speech Interface Synthesis. In *Proceedings of the 32nd International BCS Human Computer Interaction Conference (HCI '18)*. BCS Learning & Development Ltd., Swindon, UK, Article 169, 3 pages. DOI: http://dx.doi.org/10.14236/ewic/HCI2018.170

[14] Leigh Clark, Philip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew Aylett, João Cabral, Cosmin Munteanu, Justin Edwards, and Benjamin R. Cowan. 2019a. The State of Speech in HCI: Trends, Themes and Challenges. *Interacting with Computers* (09 2019). DOI: http://dx.doi.org/10.1093/iwc/iwz016

[15] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. 2019b. What Makes a Good Conversation?: Challenges in Designing Truly Conversational Agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 475, 12 pages. DOI: http://dx.doi.org/10.1145/3290605.3300705

[16] Benjamin R. Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. 2017. "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 (MobileHCI '17)*. ACM, New York, NY, USA, Article 43, 12 pages. DOI: http://dx.doi.org/10.1145/3098279.3098539

[17] Fred J. Damerau. 1964. A Technique for Computer Detection and Correction of Spelling Errors. *Commun. ACM* 7, 3 (1964), 171–176. DOI: http://dx.doi.org/10.1145/363958.363994

[18] Djellel Eddine Difallah, Michele Catasta, Gianluca Demartini, Panagiotis G. Ipeirotis, and Philippe Cudré-Mauroux. 2015. The Dynamics of Micro-Task Crowdsourcing: The Case of Amazon MTurk. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15)*. ACM, New York, NY, USA, 617–617. DOI: http://dx.doi.org/10.1145/2740908.2744109

[19] Djellel Eddine Difallah, Gianluca Demartini, and Philippe Cudré-Mauroux. 2013. Pick-a-crowd: Tell Me What You Like, and I'll Tell You What to Do. In *Proceedings of the 22nd International Conference on World Wide Web (WWW '13)*. ACM, New York, NY, USA, 367–374. DOI: http://dx.doi.org/10.1145/2488388.2488421

[20] Djellel Eddine Difallah, Gianluca Demartini, and Philippe Cudré-Mauroux. 2016. Scheduling Human Intelligence Tasks in Multi-Tenant Crowd-Powered Systems. In *Proceedings of the 25th International Conference on World Wide Web (WWW '16)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 855–865. DOI: http://dx.doi.org/10.1145/2872427.2883030

[21] Maxine Eskenazi, Gina-Anne Levow, Helen Meng, Gabriel Parent, and David Suendermann. 2013. *Crowdsourcing for speech processing: Applications to data collection, transcription and assessment*. John Wiley & Sons.

[22] Eduardo Fonseca, Jordi Pons, Xavier Favory, Frederic Font, Dmitry Bogdanov, Andrés Ferraro, Sergio Oramas, Alastair Porter, and Xavier Serra. 2017. Freesound Datasets: a platform for the creation of open audio datasets. In *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR 2017)*. Suzhou, China, 486–493.

[23] Ujwal Gadiraju, Gianluca Demartini, Ricardo Kawase, and Stefan Dietze. 2018. Crowd Anatomy Beyond the Good and Bad: Behavioral Traces for Crowd Worker Modeling and Pre-selection. *Computer Supported Cooperative Work (CSCW)* (Jun 2018). DOI: http://dx.doi.org/10.1007/s10606-018-9336-y

[24] Ujwal Gadiraju, Ricardo Kawase, and Stefan Dietze. 2014. A Taxonomy of Microtasks on the Web. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media (HT '14)*. ACM, New York, NY, USA, 218–223. DOI: http://dx.doi.org/10.1145/2631775.2631819

[25] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio Set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 776–780. DOI: http://dx.doi.org/10.1109/ICASSP.2017.7952261

[26] Jorge Goncalves, Michael Feldman, Subingqian Hu, Vassilis Kostakos, and Abraham Bernstein. 2017. Task Routing and Assignment in Crowdsourcing Based on Cognitive Abilities. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1023–1031. DOI: http://dx.doi.org/10.1145/3041021.3055128

[27] Jorge Goncalves, Denzil Ferreira, Simo Hosio, Yong Liu, Jakob Rogstadius, Hannu Kukka, and Vassilis Kostakos. 2013. Crowdsourcing on the Spot: Altruistic Use of Public Displays, Feasibility, Performance, and Behaviours. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '13)*. ACM, New York, NY, USA, 753–762. DOI: http://dx.doi.org/10.1145/2493432.2493481

[28] Jorge Goncalves, Simo Hosio, Niels van Berkel, Furqan Ahmed, and Vassilis Kostakos. 2017. CrowdPickUp: Crowdsourcing Task Pickup in the Wild. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 51 (Sept. 2017), 22 pages. DOI: http://dx.doi.org/10.1145/3130916

[29] Jorge Goncalves, Hannu Kukka, Iván Sánchez, and Vassilis Kostakos. 2016. Crowdsourcing Queue Estimations in Situ. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16)*. ACM, New York, NY, USA, 1040–1051. DOI: http://dx.doi.org/10.1145/2818048.2819997

[30] Lisa Graham. 2008. Gestalt theory in interactive media design. *Journal of Humanities & Social Sciences* 2, 1 (2008).

[31] Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P. Bigham. 2018. A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 449, 14 pages. DOI: http://dx.doi.org/10.1145/3173574.3174023

[32] Seyyed Hadi Hashemi, Kyle Williams, Ahmed El Kholy, Imed Zitouni, and Paul A. Crook. 2018. Measuring User Satisfaction on Smart Speaker Intelligent Assistants Using Intent Sensitive Query Embeddings. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18)*. ACM, New York, NY, USA, 1183–1192. DOI: http://dx.doi.org/10.1145/3269206.3271802

[33] Danula Hettiachchi, Niels van Berkel, Tilman Dingler, Fraser Allison, Vassilis Kostakos, and Jorge Goncalves. 2019a. Enabling Creative Crowd Work through Smart Speakers. In *Workshop on Designing Crowd-powered Creativity Support Systems (CHI '19 Workshop)*. 1–5. DOI:http://dx.doi.org/10.5281/zenodo.2648986

[34] Danula Hettiachchi, Niels van Berkel, Simo Hosio, Vassilis Kostakos, and Jorge Goncalves. 2019b. Effect of Cognitive Abilities on Crowdsourcing Task Performance. In *Human-Computer Interaction – INTERACT 2019*. Springer International Publishing, Cham, 442–464. DOI: http://dx.doi.org/10.1007/978-3-030-29381-9_28

[35] Simo Hosio, Jorge Goncalves, Vili Lehdonvirta, Denzil Ferreira, and Vassilis Kostakos. 2014. Situated Crowdsourcing Using a Market Model. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology (UIST '14)*. ACM, New York, NY, USA, 55–64. DOI: http://dx.doi.org/10.1145/2642918.2647362

[36] Thivya Kandappu, Nikita Jaiman, Randy Tandriansyah, Archan Misra, Shih-Fen Cheng, Cen Chen, Hoong Chuin Lau, Deepthi Chander, and Koustuv Dasgupta. 2016. TASKer: Behavioral Insights via Campus-based Experimental Mobile Crowd-sourcing. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '16)*. ACM, New York, NY, USA, 392–402. DOI:http://dx.doi.org/10.1145/2971648.2971690

[37] Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. 2012. The Face of Quality in Crowdsourcing Relevance Labels: Demographics, Personality and Labeling Accuracy. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM '12)*. ACM, New York, NY, USA, 2583–2586. DOI:http://dx.doi.org/10.1145/2396761.2398697

[38] Julia Kiseleva, Kyle Williams, Jiepu Jiang, Ahmed Hassan Awadallah, Aidan C. Crook, Imed Zitouni, and Tasos Anastasakos. 2016. Understanding User Satisfaction with Intelligent Assistants. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval (CHIIR '16)*. ACM, New York, NY, USA, 121–130. DOI: http://dx.doi.org/10.1145/2854946.2854961

[39] Ahmet Baki Kocaballi, Liliana Laranjo, and Enrico Coiera. 2019. Understanding and Measuring User Experience in Conversational Interfaces. *Interacting with Computers* 31, 2 (05 2019), 192–207. DOI: http://dx.doi.org/10.1093/iwc/iwz015

[40] Arne Köhn, Florian Stegen, and Timo Baumann. 2016. Mining the Spoken Wikipedia for Speech Data and Beyond. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (23-28). European Language Resources Association (ELRA), Paris, France.

[41] Il-Youp Kwak, Jun Ho Huh, Seung Taek Han, Iljoo Kim, and Jiwon Yoon. 2019. Voice Presentation Attack Detection Through Text-Converted Voice Command Analysis. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 598, 12 pages. DOI: http://dx.doi.org/10.1145/3290605.3300828

[42] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. 2018. Alexa, Are You Listening?: Privacy Perceptions, Concerns and Privacy-seeking Behaviors with Smart Speakers. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 102 (Nov. 2018), 31 pages. DOI:http://dx.doi.org/10.1145/3274371

[43] Adrian Leemann, Marie-José Kolly, and David Britain. 2018. The English Dialects App: The creation of a crowdsourced dialect corpus. *Ampersand* 5 (2018), 1 – 17. DOI: http://dx.doi.org/10.1016/j.amper.2017.11.001

[44] Irene Lopatovska, Katrina Rink, Ian Knight, Kieran Raines, Kevin Cosenza, Harriet Williams, Perachya Sorsche, David Hirsch, Qi Li, and Adrianna Martinez. 2019. Talk to me: Exploring user interactions with the Amazon Alexa. *Journal of Librarianship and Information Science* 51, 4 (2019), 984–997. DOI: http://dx.doi.org/10.1177/0961000618759414

[45] Michal Luria, Guy Hoffman, and Oren Zuckerman. 2017. Comparing Social Robot, Screen and Voice Interfaces for Smart-Home Control. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 580–628. DOI: http://dx.doi.org/10.1145/3025453.3025786

[46] Christine Murad, Cosmin Munteanu, Leigh Clark, and Benjamin R. Cowan. 2018. Design Guidelines for Hands-free Speech Interaction. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct (MobileHCI '18)*. ACM, New York, NY, USA, 269–276. DOI:http://dx.doi.org/10.1145/3236112.3236149

[47] Christine Murad, Cosmin Munteanu, Benjamin R. Cowan, and Leigh Clark. 2019. Revolution or Evolution? Speech Interaction and HCI Design Guidelines. *IEEE Pervasive Computing* 18, 02 (apr 2019), 33–45. DOI: http://dx.doi.org/10.1109/MPRV.2019.2906991

[48] Chelsea Myers, Anushay Furqan, Jessica Nebolsky, Karina Caro, and Jichen Zhu. 2018. Patterns for How Users Overcome Obstacles in Voice User Interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 6, 7 pages. DOI: http://dx.doi.org/10.1145/3173574.3173580

[49] Chelsea M. Myers, Anushay Furqan, and Jichen Zhu. 2019. The Impact of User Characteristics and Preferences on Performance with an Unfamiliar Voice User Interface. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 47, 9 pages. DOI:http://dx.doi.org/10.1145/3290605.3300277

[50] Christi Olson and Kelli Kemery. 2019. Voice Report: From answers to action: customer adoption of voice technology and digital assistants. (2019). https://advertiseonbing-blob.azureedge.net/blob/bingads/media/insight/whitepapers/2019/04%20apr/voice-report/bingads_2019voicereport.pdf

[51] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 527–536. DOI:http://dx.doi.org/10.18653/v1/P19-1050

[52] Aung Pyae and Paul Scifleet. 2018. Investigating Differences Between Native English and Non-native English Speakers in Interacting with a Voice User Interface: A Case of Google Home. In *Proceedings of the 30th Australian Conference on Computer-Human Interaction (OzCHI '18)*. ACM, New York, NY, USA, 548–553. DOI: http://dx.doi.org/10.1145/3292147.3292236

[53] Julie Rico. 2010. Evaluating the Social Acceptability of Multimodal Mobile Interactions. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems (CHI EA '10)*. ACM, New York, NY, USA, 2887–2890. DOI:http://dx.doi.org/10.1145/1753846.1753877

[54] Jeffrey M. Rzeszotarski and Aniket Kittur. 2011. Instrumenting the Crowd: Using Implicit Behavioral Measures to Predict Task Performance. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology (UIST '11)*. ACM, New York, NY, USA, 13–22. DOI: http://dx.doi.org/10.1145/2047196.2047199

[55] Clay Shirky. 2010. *Cognitive surplus: How technology makes consumers into collaborators*. Penguin, UK.

[56] Ben Shneiderman. 2000. The Limits of Speech Recognition. *Commun. ACM* 43, 9 (Sept. 2000), 63–65. DOI:http://dx.doi.org/10.1145/348941.348990

[57] Kinga Skorupska, Manuel Nunez, Wieslaw Kopec, and Radoslaw Nielek. 2018. Older Adults and Crowdsourcing: Android TV App for Evaluating TEDx Subtitle Quality. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 159 (Nov. 2018), 23 pages. DOI: http://dx.doi.org/10.1145/3274428

[58] David R. Thomas. 2006. A General Inductive Approach for Analyzing Qualitative Evaluation Data. *American Journal of Evaluation* 27, 2 (2006), 237–246. DOI: http://dx.doi.org/10.1177/1098214005283748

[59] Cassia Valentini-Botinhao, Xin Wang, Shinji Takaki, and Junichi Yamagishi. 2016. Speech Enhancement for a Noise-Robust Text-to-Speech Synthesis System Using Deep Recurrent Neural Networks. In *Interspeech*. 352–356. DOI: http://dx.doi.org/10.21437/Interspeech.2016-159

[60] Aditya Vashistha, Abhinav Garg, and Richard Anderson. 2019. ReCall: Crowdsourcing on Basic Phones to Financially Sustain Voice Forums. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 169, 13 pages. DOI: http://dx.doi.org/10.1145/3290605.3300399

[61] Aditya Vashistha, Pooja Sethi, and Richard Anderson. 2017. Respeak: A Voice-based, Crowd-powered Speech Transcription System. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 1855–1866. DOI:http://dx.doi.org/10.1145/3025453.3025640

[62] Aditya Vashistha, Pooja Sethi, and Richard Anderson. 2018. BSpeak: An Accessible Voice-based Crowdsourcing Marketplace for Low-Income Blind People. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 57, 13 pages. DOI: http://dx.doi.org/10.1145/3173574.3173631

[63] Yan Yan, Romer Rosales, Glenn Fung, and Jennifer G. Dy. 2011. Active Learning from Crowds. In *Proceedings of the 28th International Conference on International Conference on Machine Learning (ICML'11)*. ACM, New York, NY, USA, 1161–1168.