StreetWise: Smart Speakers vs Human Help in Public Slum Settings

Jennifer Pearson
FIT Lab, Computational Foundry
Swansea University, Swansea, UK
j.pearson@swansea.ac.uk

Simon Robinson
FIT Lab, Computational Foundry
Swansea University, Swansea, UK
s.n.w.robinson@swansea.ac.uk

Thomas Reitmaier
FIT Lab, Computational Foundry
Swansea University, Swansea, UK
thomas.reitmaier@swansea.ac.uk

Matt Jones
FIT Lab, Computational Foundry
Swansea University, Swansea, UK
matt.jones@swansea.ac.uk

Shashank Ahire
Industrial Design Centre
IIT Bombay, Mumbai, India
ahire.shashank@iitb.ac.in

Anirudha Joshi
Industrial Design Centre
IIT Bombay, Mumbai, India
anirudha@iitb.ac.in

Deepak Sahoo
FIT Lab, Computational Foundry
Swansea University, Swansea, UK
d.r.sahoo@swansea.ac.uk

Nimish Maravi
Industrial Design Centre
IIT Bombay, Mumbai, India
nimishmaravi@iiitdmj.ac.in

Bhakti Bhikne
Industrial Design Centre
IIT Bombay, Mumbai, India
bhaktibhikne@iitb.ac.in

ABSTRACT

This paper explores the use of conversational speech question and answer systems in the challenging context of public spaces in slums. A major part of this work is a comparison of the source and speed of the given responses; that is, either machine-powered and instant or human-powered and delayed. We examine these dimensions via a two-stage, multi-sited deployment. We report on a pilot deployment that helped refine the system, and a second deployment involving the installation of nine of each type of system within a large Mumbai slum for a 40-day period, resulting in over 12,000 queries. We present the findings from a detailed analysis and comparison of the two question-answer corpora; discuss how these insights might help improve machine-powered smart speakers; and, highlight the potential benefits of multi-sited public speech installations within slum environments.

CCS CONCEPTS

• **Information systems** → *Speech / audio search*; • **Human-centered computing** → Field studies; Interaction paradigms; Sound-based input / output; Interaction techniques.

KEYWORDS

Speech appliances, public space interaction, emergent users.

ACM Reference Format:
Jennifer Pearson, Simon Robinson, Thomas Reitmaier, Matt Jones, Shashank Ahire, Anirudha Joshi, Deepak Sahoo, Nimish Maravi, and Bhakti Bhikne. 2019. StreetWise: Smart Speakers vs Human Help in Public Slum Settings. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland, UK*. ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3290605.3300326

CHI 2019, May 4–9, 2019, Glasgow, Scotland, UK

ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300326

1 INTRODUCTION

Human-driven question and answer (Q&A) systems have been deployed for many years. Consider, for example, text-based approaches in online discussion boards [18], local area app-based forums such as Neighbourly [6] as well as audio-based services such as Question Box [17], or interactive voice response (IVR) systems such as the Spoken Web [10]. Such approaches have also been used to complement machine curated information – consider, for instance, Google Maps' requests to its users of *"Someone has a question about…"*, which highlight the need for a richer, more personal response than an algorithm can currently provide alone.

Despite the clear value of human responses, the deployment of automated data driven systems is increasing. In particular, speech-based services have proliferated in recent years, and are now widely used in the form of mobile services such as Siri; and, more saliently for our work, as home-based smart speakers such as Amazon's *Echo* or Google's *Home*. Prior research has shown the range of questions users pose to speech services [13], and has also highlighted the many potential issues associated with AI and speech recognition [1].

In the work reported here we explore the differences in terms of efficacy, use and adoption between a speech system powered by humans and one powered by AI. Our motivation is both to look at human alternatives to the assumed primacy of AI speech systems; and, to consider ways of improving current AI systems from insights gathered. In addition, our work attempts for the first time to provide a longitudinal evaluation of such systems in public settings, as opposed to the designed-for norm of personal or home use.

In previous work we have demonstrated the potential for smart speakers in public places, testing this via a short-term Wizard-of-Oz probe [21]. Here, we deploy two fully-functional public space speech systems, longitudinally, in the challenging setting of Dharavi – a large slum in Mumbai, India. Speech interaction has long been seen as a particularly beneficial modality in contexts such as Dharavi, due to the relatively high rates of textual and technological illiteracy. Such communities, however, are typically resource constrained relative to the "traditional" users that current smart speakers target. Apart from constraints related to educational attainment, other issues often include an inability to afford such devices, a lack of the necessary home wireless and power infrastructure; and, highly limited personal living space. These obstacles put conversational speech systems out of reach for many Dharavi residents. The installation of such systems in public settings, then, provides a wider reach, and potentially gives the opportunity to access information to which such users may otherwise have been denied.

In the rest of this paper we describe several design iterations of public space interactive speech systems, and the pilot and full deployment of 18 (nine AI-based; nine human-powered) of these in public slum settings for a 40-day period. We describe the main findings from over 12,000 questions received, highlighting similarities and differences between AI and human-curated approaches, as well as discussing how these findings could be used more widely to improve current speech-based services. The work sheds new light on the trade-offs between human- and machine-powered systems in challenging public environments, and provides a foundation for future work in this area.

2 RELATED WORK

To contextualise our work, we consider existing speech Q&A systems along two primary dimensions, as shown in Fig. 1: first, whether the response is *instantaneous* or *delayed*, and second, whether answers are provided through *human* understanding or *machine* learning. We incorporate a third, colour-coded dimension to indicate through which *types* of devices, and in which *settings* speech systems are typically accessed, namely on *personal mobile* devices (blue), *shared* devices situated in the *home* (green) and *shared* devices installed in *public* settings (red).

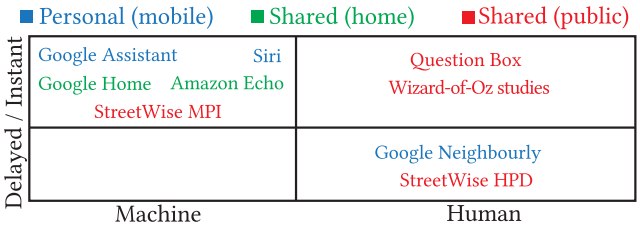

Figure 1: Speed vs. respondent speech Q&A system matrix.

Advances in machine learning, cloud computing and natural language processing have led to a proliferation of voice user interfaces and interactions in the upper left quadrant of Fig. 1: from voice-controlled conversational agents embedded into our mobile phones—e.g., Google's *Assistant* and Apple's *Siri*—to smart speakers, a new class of dedicated speech devices that situate voice user interfaces inside the home. While we are fascinated by the new opportunities this modality enables, here we primarily draw upon critical scholarship within HCI to better appreciate the *frictions* and *limitations* of such systems to inform the design of our AI-powered prototype. Following this, we discuss examples of systems that rely on human help, rather than human intelligence. We present these alongside scholarship which unpacks different qualities and capabilities between human understanding and artificial intelligence to better understand how we can learn from human-powered speech systems to inform the design of future conversational agents.

Conversational agents

A recurring theme of research into conversational agents (CAs) is that *"user expectations of CA systems remain far from the practical realities of use"* [11]. Precisely because of this gap, researchers take issue with calling conversational systems *conversational* [11, 15]. Consider participants in Luger and Sellen's study of CA users, who *"described making use of a particular economy of language. Dropping words other than 'keywords', removing colloquial or complex words, reducing the number of words used, using more specific terms, altering enunciation, speaking more slowly/clearly and changing accent were the most commonly described tactics"* [11].

In that study, the success and satisfaction users reported with mobile CAs depended in large part on the extent to which they developed a mental model of system capability and intelligence. For instance, users who drew less from such a technical frame and more upon a model of human-human dialogue, such as by *"beginning their interactions with fuller and more natural sentences"* [11], became increasingly frustrated with the device and tended to blame themselves when interactions failed. So *"the principle [sic] use-case of the CA was 'hands-free', meaning that an alternative primary task, rather than the conversation, was the focus of attention"* [11].

Smart speakers

According to surveys in 2018, 18 % of American adults [13] and 10 % of Britons [28] own a smart speaker, and next to listening to music and podcasts, respondents primarily use them to seek answers to general questions. Porcheron et al.'s ethnomethodological study of smart speaker use [15] draws attention to the work required to situate the Amazon Echo smart speaker into the complex social context of the home. That is, into a multi-activity context where potentially multiple people are interacting with the device (and with each other) simultaneously. The research lays bare how poorly the device fits into multi-party conversations; for instance when family members interacted with the smart speaker during a meal while others were conversing concurrently. Given this mismatch, Porcheron et al. suggest that future work on voice user interfaces (VUIs) should, for the time being, shift *"from conversation design to [...] request/response design"* [15].

We are also mindful that studies of smart speaker and CA use are predominantly located [23] in Euro-American contexts [2, 11, 15]. Even for native English speakers, research reports how speech systems struggle to correctly recognise words when users speak with accents [12]. Furthermore, in interviews of smart speaker users, Pyae and Joelsson found some users lamenting that *"non-English words"*, such as place names, *"are not correctly captured by the device"* [16]. Given such numerous accounts of interactional and practical limitations, we next consider how speech systems that leverage human *understanding*, rather than artificial intelligence, can overcome and help us better understand these issues.

Human help

A fundamental limitation of data-driven systems, of course, is that they can only ever be as good as the data they derive their insights from. Simply put, data is sparser outside of the major centres of industrial research and development in the hyperdeveloped world [24]. It is no surprise, then, that a company like Google is trialling a system called *Neighbourly* [6] after announcing Hindi language support for their Assistant platform [5]. Neighbourly is an Android app that allows users to type or speak a question. If spoken, speech-to-text algorithms then transcribe the question in eight Indian languages. Once a question is posed, nearby users can answer the question by speaking or typing their answers into the system, which are then automatically transcribed if necessary. Neighbourly is therefore a prime example of a human-powered delayed Q&A system. It and companies such as Uliza[1] demonstrate "business cases" for human-powered Q&A, and provide additional motivations for this work.

Robinson et al.'s Wizard-of-Oz system [21] replaces the AI back-end with a human. When activated, the user is connected with a remote operator through the 'smart speaker' interface. Such a system can facilitate multi-lingual input/output with the potential for richer, more contextual and even personal responses. Question Box [17] is another example of a human-powered instant speech Q&A system, and has been effectively used for specific domain areas such as antenatal care or agricultural extension services. While these human-powered instant-response approaches might be considered an 'ideal' solution, we are wary that such one-to-one human-driven systems are difficult to scale, however.

3 INITIAL DESIGN CONSIDERATIONS

Our aim at the outset of this work was to understand options and opportunities for speech based Q&A systems in resource constrained settings among 'emergent' users of technology: those only recently getting access to more advanced digital devices and services, but who consequently don't have the same technological abilities or media literacies. Devanuj and Joshi [3] characterise emergent users in India based on their abilities to use ICTs, namely as: basic users, navigators, inputters, savers, account holders, and transactors. For this research we are particularly interested in the first three categories. Basic users are those who are only able to use products with static buttons, such as those found on a featurephone. Navigators are able to abstract away concepts, which allows hierarchical navigation [27], and enables them to use smartphones or navigate websites. Text inputters are able to type, and do many more tasks, such as searching [14]. With this research we seek to target and blur the text inputters category. Text input in Indian languages has been reported to be challenging [9], which helps explain why this stood out as a distinct emergent user category in India. However, if speech based systems were to take away the need to type (and read) text, would more people (including navigators and basic users) be able to search for information? What types of things would they ask; and, would an AI-based system be able to understand and answer them?

Reflecting on the comparative advantages and disadvantages of existing human and machine powered approaches, and steered by the design space and guidelines of [21], we opted to prototype and deploy two systems, evaluating the immediate vs. delayed dimension of that design space:

Machine-powered-instant (MPI): A user asks a question and instantly receives an AI response.

Human-powered-delayed (HPD): A user asks a question, instantly receives a numeric reference code, and returns after a delay (up to 10 minutes) to retrieve a human response.

The overall goal of this work, then, was to explore, through a longitudinal deployment of the two approaches:

1. Insights into the potential use-cases and benefits of public-space speech systems for emergent users;
2. How MPI and HPD systems cope in noisy environments, with a range of accents and multi-lingual users;

3. Potential approaches to support distributed, multi-site public speech interaction;
4. Differences and benefits of delayed human-curated responses against real-time machine-powered responses;
5. How HPD systems might address shortcomings of MPI systems, informing the design of future CAs.

4 PROTOTYPES: PILOT DEPLOYMENTS

Before attempting to longitudinally deploy multiple prototypes, we first chose to build a single version of each system to test in lab-studies and a pilot deployment within Dharavi. The reason for this was to assess the usability of the front-end systems in Dharavi's challenging environment (even more so with no training or support for users); and, stress-test back-end services to verify the feasibility of the approach.

Figure 2 shows the initial prototypes we designed. Both consisted of a speaker, microphone and button; and, in the HPD case, a receipt printer. Both also used mobile data-connections and were battery powered to provide flexibility in placement location. The MPI system was adapted from an off-the-shelf AIY Voice Kit [8]. Pressing its button causes it to start listening for spoken input. When a pause in speech is detected, the Google Assistant API processes and returns a spoken response immediately (i.e., within a few seconds).

The HPD system, while similar in external form, works very differently. Pressing its button initiates listening for speech, but when a pause is detected, audio is sent to an API powered by a voice crowdsourcing partner company[1]. In parallel, a ticket is printed (via the slot at the base of Fig. 2, right), giving the questioner a number to use to retrieve their answer. Over the next few minutes (targeting a maximum of 10 min), a local-language-speaking answerer listens to the audio file and records an audio response. The questioner can later return, press the button again and speak the ticket number into the microphone to retrieve their answer. Throughout this interaction, the questioner is prompted in Hindi via spoken instructions to *"speak now"*, *"read the number on the ticket"*, etc. Answers are not linked to the specific machine that submitted the question, so a deployment of multiple systems would allow questioners to ask in one location and retrieve their answer in another.

Lab testing pilot

We recruited 13 emergent users (11M;2F) for a short pilot study to help test and improve the systems before deployment. All participants were housekeepers or cleaners with a primary school level of literacy in their local language (Hindi or Marathi). In an informal, workshop-like setting, we asked participants to pose questions to each of the systems, and observed their interactions and any challenges.

[1]Uliza – see: https://www.uliza.org/

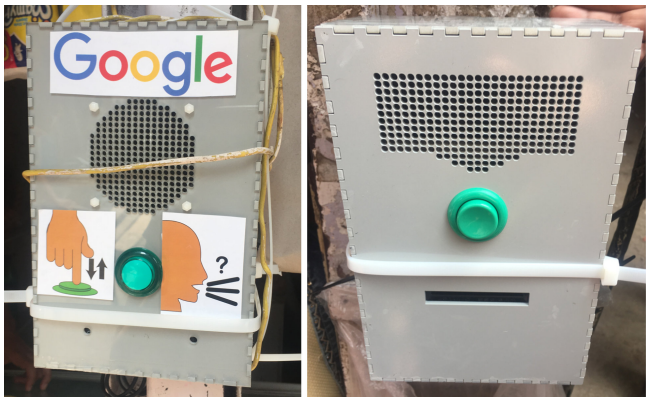

Figure 2: Initial prototypes. Left: the machine-powered-instant system (MPI), showing pictogram signage added after the lab study. Right: the human-powered-delayed system (HPD) in its original form.

This pilot identified a range of improvements, such as a suggestion to add pictogram step-by-step instructions (see Fig. 2, left), and the need to highlight the button to ensure that it was not confused for other aspects such as the speaker grille or fixings. On the advice of local researchers, we also added the Google logo to both boxes as it was felt to be a recognisable symbol for local residents in Dharavi (regardless of whether residents had used an internet search engine themselves), and might give a clue as to the system's purpose.

Deployment pilot

We installed one of each of the refined, lab-tested systems in two separate locations under the supervision of 'caretakers'. These were owners of local businesses (a dairy and an ice cream shop) whom we recruited to take responsibility for the devices, protecting them from the weather and preventing theft, as well as keeping batteries charged, and ensuring the systems were placed in suitable places to be used by passersby. Once installation was complete, we left each box in-situ for a period of four days, during which we observed each system's usage from a distance, and tracked usage via remote interaction logs.

A range of issues were encountered in the deployment pilot. Most of these challenges were technical in their nature: for example, unreliable 4G signal led to frequent connection failures, which were frustrating for users. Similarly, battery power, while providing flexibility in location, became a problem when caretakers forgot to charge the devices.

With the MPI system there were a range of API-specific issues, such as quota limits and periodic backend failures. For the HPD system, the problems were more disruptive. One of the key problems was the ticket printer which, due to the humid and dusty deployment environment, regularly became

jammed. While the audio feedback that was incorporated throughout the process mitigated this problem (ticket numbers were also spoken aloud), it was clear that we needed to extend the coverage of these system state messages to expose system state more helpfully. For example, any errors when sending or receiving answers (caused by the unreliable network coverage mentioned above) led to user confusion, which was exacerbated more on the HPD device due to the delayed aspect of the system. Finally, recognition of spoken ticket numbers was unreliable in Dharavi's noisy environment, and caused many of these inputs to be mistakenly submitted as questions.

Discussion

Following the pilots, we redesigned both systems with a focus on addressing the interaction shortcomings of the initial designs. First, we improved the labelling and layout of both systems' boxes, separating their components into a linear flow with clear pictographic and simple written instructions (see Fig. 3). The instructions did not indicate whether either system was powered by a machine or a human, but simply invited passersby to *"ask me any question"* (Hindi). We also switched to mains power after caretakers advised us that grid availability was better than expected. We tested a range of 4G network providers in each deployment location, and selected the ones with the best coverage. Finally, we switched from a standard button to one able to light up or blink.

With the HPD system in particular we made several larger changes, removing the printer entirely and replacing this with an OLED display and keypad to input and output question numbers. Initially we considered using a 7-segment display, but opted for a more legible alternatives [25] for outdoor deployment. This change improved the design both by removing the potential for paper jams or misrecognised spoken numbers, and also by allowing the use of the OLED screen to provide visual feedback of the system's state (for example, when downloading responses, or as an indicator of when speech was being recorded).

5 LONGITUDINAL DEPLOYMENT

After refining the design of the two systems, we returned to Dharavi for a second deployment. In order to investigate the potential benefits of having multiple devices in the same local vicinity over longer periods of time, we installed nine MPI and nine HPD systems in various locations spread over an area of approximately 2 km^2 within Dharavi (see Fig. 4). We selected the deployment locations carefully in order to ensure that the systems were all available to broadly the same population, but also to keep some distance between user groups. This was to encourage use of more than a single system in the same local vicinity, while reducing the potential confusion of encountering two installations of different

Figure 3: The revised speech systems pictured in-situ in Dharavi during the longitudinal deployment. Left: the deployed MPI system; right: the deployed HPD system.

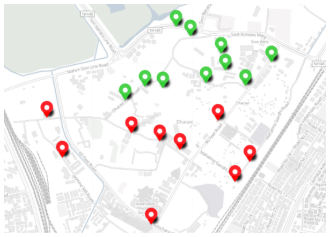

Figure 4: MPI and HPD system locations in Dharavi.

Red markers indicate machine-powered-instant systems; green markers indicate human-powered-delayed systems.

types. In the same approach as the pilot deployment, we recruited local business owners to take care of the devices, keeping them powered, dry and secure. We selected a range of different business types including stationers, bike mechanics, dairies, phone repair stores, greengrocers and chai stands, and compensated each caretaker with ₹5000. We instructed the caretakers to leave the systems in a public facing area—connected to mains power at all times—during normal business hours (typically around 10am to 10pm, with an hour or two break during mid-afternoon) and deployed for a period of 40 days in total.[2].

Data capture, sampling and analysis

We archived the audio files of all interactions and responses with each system, as well as the times of day the boxes were online, for later analysis. These files were first sampled, then used in a tagging process that allowed us to analyse patterns of use across the systems. As part of the investigation into the differences between the systems, over the course of the deployment we also fed 100 randomly-sampled MPI recordings into the HPD system to assess differences between responses to identical questions.

Over the 40-day deployment, we recorded a total of 12,158 interactions across the 18 installations, consisting of 8,174 on

[2]For logistical reasons, MPI systems were installed four days later than HPD systems, but were removed from service at the end of the study four days later than HPD systems, so both types were in action for a total of 40 days

the nine MPI systems, and 3,984 on the nine HPD systems. We used stratified sampling to select 40 % of interactions (from random boxes and times) from both the MPI (totalling 3,270 of 8,174) and the HPD (totalling 1,595 of 3,984) systems to analyse in more detail, following the methodology from previous audio-related emergent user research [26].

While we do initially describe the raw data in the form of absolute numbers, we also wanted to make direct comparisons between the *proportion* of interactions of particular types between systems. However, due to the significant differences in the quantity of interactions on the HPD system versus the MPI system (i.e., more than double), we opted to normalise the data, describing each topic, response or tag, etc., as a percentage of the total analysed interactions.

The first stage of analysis was to transcribe the local-language audio files to English text[3] using a combination of automatic and human-curated translation and transcription. In the case of the HPD system, all sampled queries and responses were translated and transcribed by native speakers at the point of answering. All MPI recordings were initially automatically transcribed and translated to English using Google's translation API. While this process easily identifies blank recordings and simple queries, it often fails to properly capture more complex sentences, thus around 18 % of recordings required further processing by human translators.

The next stage in the process involved manually analysing each interaction to categorise and tag as detailed below. To ensure consistency, a single researcher (native English speaker) processed all translated interactions and answers. First, they listened to the recording and tagged the language and, where possible, the type of speaker (male, female, child or unknown). They then studied the transcript and classified the type of interaction using the categories in [21] as a starting point. That is, was the interaction a question (basic facts, contextual information, philosophical questions, system-directed questions or other), not a question (a statement or request), or other noise/blank/unintelligible.

Following this, the researcher then tagged each interaction to help categorise them during analysis. These tags helped identify patterns in the type of questions asked to each system. For example, a question asking *"what is the capital of Saudi Arabia"* would be tagged with 'what', 'location', 'country' and 'geography'; a question asking *"when did India gain its independence?"* would be tagged with 'when', 'date', 'location', 'country', and 'history'. Further tags were built up over the course of the tagging process to indicate patterns in the data. For example, a question such as *"when will I fall in love?"* would be tagged with 'prediction', 'personal' and 'when' to indicate it is a query about the future of one's self.

[3]While we are aware that this process adds complexity, we were constrained to perform this action as English is the team's only common language.

Table 1: Query categories for each system, showing the percentage of valid interactions in each category. Interactions are categorised as either a *question* or *not a question,* and then into further sub-categories as detailed in Section 5.

	MPI (%)	HPD (%)
Question:	**85.7**	**96.3**
Basic fact	50.1	58.2
Contextual queries	18.6	24.5
Questions directed at machine	11.7	3.6
Philosophical question	5.4	9.9
Not a question:	**14.3**	**3.7**
Statement not requiring a response	8.2	2.8
Request	6.1	1.0

Finally, the response provided was cross-referenced with the question to determine if the answer was either:

- Blank, unable to answer or *"I don't know"*;
- Relevant to the interaction; that is, if it either answers a question correctly, or if it correctly relates to an unanswerable interaction. For example, a question asking *"when will I become a millionaire?"* can either be answered formally (e.g., *"I don't know"*); or, it can be answered creatively, for instance with a relevant response such as *"If you work hard at what you do, you can become a millionaire sooner than you think"*[4];
- Not relevant to the original interaction, but not blank or *"I don't know"*, etc. For instance, if a question has been misheard or an error has occurred, e.g., a question of *"how many planets are there?"* being answered with *"In India there are 28 states and 7 union territories"*[4].

6 RESULTS

Figure 5 summarises all interactions for each system over time.[2] As can be seen from the chart, there is an initial period of novelty effect at the start of both deployments, but overall the average number of interactions per day settles down after around one week. Figure 5 also shows the millimetres of rainfall per day, a factor we were intrigued to investigate in relation to both the times boxes were online and the number of questions asked.

Overall there was a large spread in the quantity of query interactions per installed system, ranging from 117–836 per box for the HPD system, and from 121–1968 per box for the MPI system. We attribute this large dispersion in the number of questions to many factors including the type of business (i.e., high vs low foot-traffic), the number of hours a day the system was turned on (i.e., due to the opening times of the business, or forgetful caretakers) and the enthusiasm of the caretakers themselves (some were keener than others,

[4]These are actual Q&A pairs received during the deployment

with many encouraging use, but with others actively discouraging excessive interaction). On average, however, each HPD installation could expect to receive 11 questions per day, and each MPI installation 21 questions per day. Figure 6 illustrates the time of day interactions took place, indicating that the HPD system tended to be more popular earlier in the day, whereas the MPI system was more popular later at night. These figures correlate to the times of day the boxes were online, and could be explained by the geographical locations of each type of box (HPDs were placed in the north of Dharavi where there are more leather shops, while MPIs were placed in the south where Kumbharwada (potters' village) is located).

Turning now to the interactions themselves. Across both data sets, we categorised 1,328 (27 %) of the 4,865 sampled interactions as either blank (436), unintelligible (213), incomplete (110), background conversation (459) or system bugs (110). This left a total of 3,535 valid interactions across both systems; 2,519 MPI and 1,016 HPD. As we anticipated, given the locations of the systems, there was a lot of noise in the recordings, including background conversation, heavy rain and vehicle horns. In many of the quieter HPD recordings, we were also able to identify typing on the keypad, indicating that returning users sometimes pressed the button before attempting to enter their 4-digit code.

Overall, the languages used for questions were very similar across both systems: typically questions were asked in Hindi or English, but also commonly in a hybrid combination of Hindi with a mixture of Marathi, English and Urdu words. Interestingly, the demographics of identifiable[5] users of the systems were are dominated by adult males (59 %) and children (38 %), with adult females accounting for just 3 % of the analysed interactions. These findings replicate gender imbalances of IVR systems, such as Sangeet Swara [26] and Polly [19, 20], which were both primarily used by men.

Topics of conversation

Table 1 summarises the normalised interaction categories for all *valid* interactions. In general, the types of query were similar across systems: questions were more popular than statements in both cases, but 10 % more popular on the human-powered device than the machine-powered one. Questions categorised as basic facts, contextual and philosophical queries were 8 %, 6 % and 4.5 % more popular on the HPD than the MPI, respectively. Questions directed at the machine itself, however, (for example, *"who are you?"*, *"who made you?"*, etc.) were 8 % more popular on the MPI system.

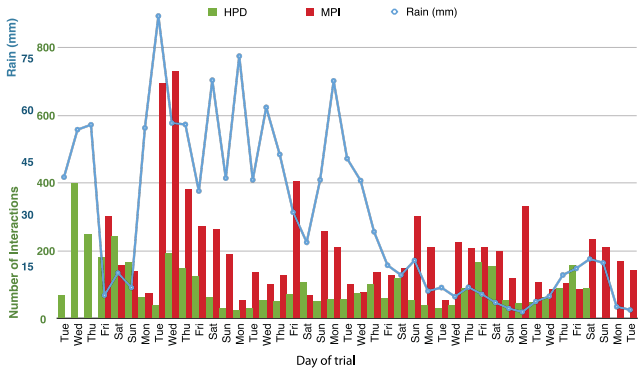

Figure 5: Usage over time, showing the total (absolute) number of interactions received from the HPD (green) and MPI (red) systems over the 40-day deployment. Also shown here (blue) is the amount of rainfall in Dharavi on each day.

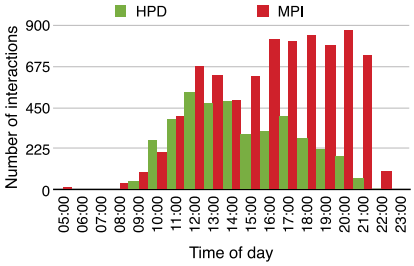

Figure 6: HPD (green) and MPI (red) interactions during each hour of the day. The majority of caretaker businesses opened from 10am–10pm.

While these broad categorisations are useful to determine the general types of interaction, we also wanted to look in more detail at the types of query on each system. Using the extensive database of textual tags we added to each interaction, we were then able to categorise the queries with a finer level of granularity. Table 2 shows the most prominent query subjects across both systems[6] (shown as a percentage of the total valid interactions for that system). Highly prominent subjects were queries relating to Geography, Politics, Music (both requests for and questions about), History and Science.

Table 3 illustrates some of the more interesting category tags. In general, local information (e.g., places, weather, commodity prices etc.) were particularly popular in both cases (MPI: 14.3 % and HPD: 20.9 %). Personal questions about one's self or a known person, (e.g., *"Where is my grandmother's town?"* (MPI), *"[name] who works in [place] is honest or thief?"* (HPD) or *"who stole my meter cable?"* (HPD)) were also popular, in particular in the HPD system which saw 8.2 % tagged as such; almost double that of the MPI system.

Many query interactions were categorised as predictions (MPI: 2.5 % and HPD: 7 %). Personal predictions accounted

[5]For the most part, we were able to identify whether the interaction was made by a male, female or child by listening to the audio recording. Those that we were not certain of were marked as unidentifiable (a total of 183 of the total valid interactions). In some cases, the interaction included multiple users (i.e., one person begins a question and another finishes it); these occasions are included in the demographic numbers stated.

[6]Note: unlike the category classification shown in Table 1, tags are *not* exclusive, so an interaction could be tagged with more than one subject.

Table 2: The most popular query tags on each system by subject, shown as a percentage of total valid interactions.

	MPI (%)	HPD (%)		MPI (%)	HPD (%)
Geography	15.0	7.3	Politics	9.7	7.4
Music	4.4	1.2	History	3.7	7.3
Science	2.9	8.7	Travel	3.7	4.6
Sport	2.2	3.1	Education	1.9	1.5
Health	1.9	4.9	Music	4.3	1.2
Technology	1.4	6.4	Religion	1.8	1.2
Maths	1.0	2.4	Economics	1.0	2.3

Table 3: Additional query tags, given as a percentage of total valid interactions.

	MPI (%)	HPD (%)		MPI (%)	HPD (%)
Local	14.3	20.9	Search term	8.7	2.6
Personal	4.3	8.2	Prediction	2.5	7.0
Subjective	0.7	1.9	Advice	1.7	4.0
News	0.6	2.0	The time	2.7	0.6

for the majority of these queries, (e.g., *"what is my destiny?"* (MPI) or *"will I pass in 10th standard?"* (HPD)). Political predictions (e.g., *"how long the Modi government will last?"* (HPD)), economical predictions (e.g., *"when do we kids will have a better future?"* (HPD)), and sporting outcomes (e.g., *"who is going to win in the Brazil and Mexico matches?"* (MPI)) were also included in these tags.

There were also many queries categorised as asking for advice, which occurred over twice as often in the HPD system than the MPI (4 % versus 1.7 %). For example, *"which is the ideal road side business in Mumbai?"* (HPD), *"what should I do to impress a girl?"* (MPI), or *"what to do to get rid of my father's alcohol habit?"* (HPD). Questions tagged as subjective (e.g., *"which is the best footballer in the world?"* (HPD), *"do aliens exist?"* (HPD) or *"who is the alpha between the girl or boy?"* (MPI)) were also over twice as popular on the HPD system than the MPI system (1.9 % versus 0.7 %).

"What time is it?" was the single most asked question on the MPI system, making up 2.7 % of all valid interactions. This question also occurred, albeit less often, on the HPD system with 0.6 % of interactions. News was a popular tag, in particular in the HPD system, with 2 % of queries, compared to 0.6 % on the MPI system. During the deployment there was a bridge collapse in Mumbai, and we received several queries relating to this. For example, *"how many bridges in Mumbai are on the verge of breaking?"* (HPD), *"is Andheri road active?"* (HPD) and *"what happened in Andheri?"* (MPI). Similarly, a plastic ban was enforced during the deployment,

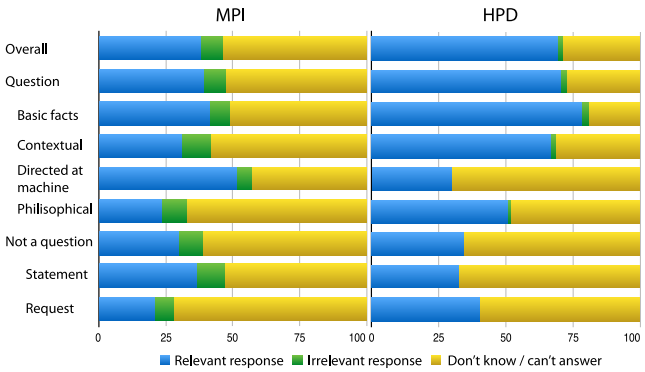

Figure 7: Distribution of responses to all *valid* queries.

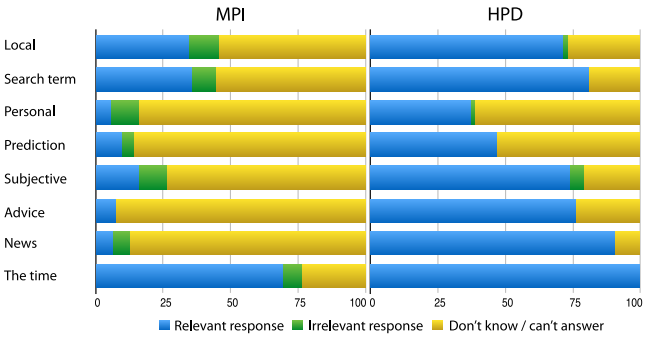

Figure 8: Distribution of responses for selected question tags.

and hence we received several questions relating to this; for example, *"why did Modi stop plastic bags?"* (MPI) and *"since when is plastic banned in Maharashtra?"* (HPD).

Ability of systems to answer questions

Figure 7 illustrates the overall and category break down of responses to all *valid* interactions on each system. Shown as a percentage, the blue areas indicate relevant responses, green are irrelevant responses and yellow are blank, unable to answer or *"I don't know"* responses (see Section 5)

Overall, the HPD system produced more relevant responses to queries (HPD: 69.1 % versus MPI: 37.9 %) and very few irrelevant responses (2.1 %) compared to the MPI system (8.4 %). A Chi-squared test for independence on this data shows it to be highly significant (χ^2 = 290.2, df=2, p < 0.00001). In terms of the categories of query received, the HPD system produced significantly more relevant responses in all *question* cases, apart from questions directed at the machine itself, where the MPI version responded with a relevant answer 21.6 % more often (χ^2 = 10.7, df=2, p < 0.005) (see Fig. 7). Interactions categorised as *not a question* were answered similarly in both cases and were not found to be statistically different overall (χ^2 = 3.8, df=2, p = 0.15) nor for statements (χ^2 = 4.1, df=2, p = 0.13) or requests (χ^2 = 2.5, df=2, p = 0.29).

Turning now to interesting interaction tags. As Fig. 8 shows, the human-powered system responded with relevant answers to local questions twice as often as the machine powered system (71 % versus 34 %). This was sometimes due to the MPI back-end mishearing a query, but also because this type of information is not always easy to locate via a simple search. One question, for instance, which was originally asked to an MPI box but was one of the randomly selected 100 questions fed into the HPD system, asks about a local business: *"I am Dharavi, where is sony fast food?"*. While the MPI system responded with an *"I can't help you"* retort, the HPD answerer looked up the location of the local business and gave appropriate directions.

With regards to personal queries (e.g., *"what is my father's name"*), while both systems struggled to respond appropriately to even half of such queries, the HPD system did perform better than the MPI in providing relevant responses to personal questions (HPD: 37.3 % versus MPI: 5.6 %). Responses to interactions tagged as subjective and advice were both significantly more relevantly answered on the HPD system than the MPI, presumably due to the creativity required for such queries.

Interactions tagged as predictions were answered with relevant responses 46.5 % of the time on the HPD system, compared with only 9.4 % of the time on the MPI system (χ^2 = 24.5, df=2, p < 0.00001). Similar queries were asked to both types of systems but answered in different ways. For example, one prediction query across both systems was *"who will be the next Prime Minister?"*. This exact question was asked 12 times over both systems (7 HPD, 5 MPI). Of these, five of the seven HPD responses were relevant (e.g., *"in future we don't know who will become the Prime Minister – maybe you will become the Prime Minister in 2019, but right now the Prime Minister is Narendra Modi"*), while all five MPI responses were categorised as unable to answer.

Richness of conversation

The majority of interactions with both systems were posed as full sentences, however, in some cases, questions were posed more like a search engine query, (e.g., *"India capital"* (HPD) or *"parts of body"* (MPI)). These were far more prominent on the MPI system (MPI: 8.7 % versus HPD: 2.6 %). The second entry in Fig. 8 illustrates the responses provided for these queries, and shows a strong success rate for relevant responses on the HPD system (80.8 %) versus the MPI system (35.5 %). What this does not illustrate, however, is the percentage of relevant answers of the inverse type of query (i.e., complete sentences). This percentage is actually significantly lower than search terms for both systems (HPD: 43.4 %, MPI: 28.8 %). Both systems performed better, therefore, when the interactions were phrased as search terms as opposed to using complete sentences. Interestingly, the difference was far bigger on the HPD system (43.4 % for complete sentences, 80.8 % for search terms) than on the MPI system (28.8 % for complete sentences, 35.5 % for search terms).

There were instances where multiple people formed part of the same overall interaction, sometimes repeating the same question, often because one person was giving an example query. The HPD system coped well with repeated interactions, providing 80.6 % relevant responses, but the MPI system struggled with such repetition; in contrast it provided only 14.3 % relevant responses. In other cases, multiple people can be heard contributing to parts of the same question. For instance, P1: *"which water body supplies water to Mumbai"*; P2: *"and what is its position right now, how full is it?"* (HPD). Furthermore, the MPI system struggled to answer two questions at once regardless of how many people were part of the interaction. For example, *"how is the weather and what is the time in New York?"* and *"the time and tide"* both received *"I don't know"* responses. In fact, the MPI system did not give a single relevant response to any interactions with multiple queries, while the HPD system answered with a relevant and complete response (i.e., answering all of the query) 68.2 % of the time.

There were instances, particularly on the MPI system, where people continued to interact with the system multiple times consecutively. One girl, for instance, chatted to an MPI box for over an hour, with a total of 135 interactions, making up almost half of all interactions for that particular system installation. The type of interactions she had with the machine began with asking about weather, with many follow-up questions and statements asking why the system did not understand. She continued chatting by asking if it could dance or sing, requesting music, inquiring about its family and even asking *"can you be my friend?"*. After many *"I don't know"* answers over 72 min she ended with several angry statement interactions such as *"I am rejecting you"* and *"I can kill you because you do not respond properly"*.

Many questions were phrased in a way that machines currently struggle with, but that humans can answer easily. For example, *"do they really cook and eat a dogs in China?"* was answered relevantly by the HPD system during the trial but when fed into the MPI system for comparison, it was incorrectly interpreted and consequently gave an *"I don't know"* response. Similarly, the question *"I want to secure a labour license for painting – where will I have to go for it?"* (HPD) was also answered relevantly by the HPD system, but was not interpreted correctly or answered properly after being fed into the MPI system.

Occasionally we overheard people talking about the system, in particular in the HPD where we can hear people explaining to others about the 4-digit codes used (e.g., P1: *"what is history?"* P2: *"now here comes a number, note down that number, now come after 15 minutes"*. In one case, a

shopkeeper was overheard encouraging usage by offering a discount on his goods, *"ask something and twenty rupees will be deducted – ask something!"*. In another interaction, others can be heard in the background saying to ask something they do not know the answer to: P1: *"when was Gandhiji born?"*; P2: *"ask something that you and I don't know"* P3: *"everyone knows when Gandhiji was born"* (HPD). The HPD system also overheard several people pressing the button and attempting to speak the 4-digit code, but, far more commonly, we can hear people attempting to enter the code into the keypad. We believe this to be the main reason for the HPD system having more invalid (blank) interactions than the MPI system.

Philosophical questions were more prominent in the HPD system (9.9 % versus 5.4 %), and relevant responses were also more likely (HPD: 50.5 % versus MPI: 23.4 % (χ^2 = 22.825, df=2, p < 0.0001). HPD systems also tended to obtain richer more meaningful questions as well as more creative and relevant responses. For instance, *"there are so many atrocities happening on the girls in India, why do not the police take action quickly?"* (HPD) was answered relevantly and richly with: *"yes, we believe that there is a lot of torture happening on the girls in India, but it happens because nobody reports them"*. Feeding this audio file into the MPI system resulted in a correct speech-to-text conversion, but the response given was *"I don't know"*. Another example asks *"when will we improve on the days of the poor?"* (HPD), which was answered as *"the Government of India has organised several events, like those who are successful, will improve on the days of all the people"* by the HPD system, but once again, when fed into the MPI system not only did it provide an *"I don't know"* response, but it also did not properly transcribe the question.

In comparison, there were very few particularly rich and meaningful questions asked to the MPI system. Those that were, for instance, *"why does an honest person get bad-luck?"* (MPI), were typically answered by an *"I don't know"* response. This result suggests that users got to know the systems over time and adjusted their question asking accordingly. Learning that a human is answering could entice users to ask more meaningful questions, while knowing that a machine will usually say *"I don't know"* may put them off asking such questions in the future. Note that the installations were not labelled with the type of response (i.e., HPD vs. MPI), so questioners had to discover this behaviour through actual usage. The back-end of the MPI system does have standard responses for certain, presumably common philosophical questions. For instance, the typical *"which came first, the chicken or egg?"* question was asked 22 times across both systems, and was answered relevantly in most cases. In fact, this one question alone accounted for 22 % of all relevant MPI responses to philosophical questions.

Monsoon season in Mumbai presents many challenges for inhabitants, and our deployment. Figure 5 illustrates the rainfall recorded during the trial, but seems to show little in the way of correlation. However, while in Dharavi, we did notice several issues and usage observations related to the rainy season. First, the noise of torrential rain causes issues for speech recognition, which may result in increased questions after receiving a lower than usual recognition rate. Most of the businesses we recruited to house the systems were small, often with little more than a tarpaulin to shelter them from the elements. The issue here is that the devices were seen by many caretakers as expensive and fragile, and as a result were often taken offline for safe-keeping during extreme weather. On the other hand, however, the rain often forces passersby to take shelter in local businesses, bringing them closer to any boxes still online.

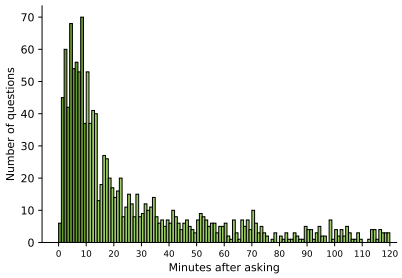

Figure 9: HPD system interaction frequency based on the amount of time elapsed between asking a question and retrieving the answer (minutes).

HPD answer retrieval

Of the 3,884 in-situ queries (e.g., the total without the 100 MPI recordings we submitted cross-system), 1,742 (44.9 %) answers were retrieved. While on the surface this rate seems relatively low, we must also take into account that there was a very high proportion of non-valid (accidental) query interactions. Assuming a similar ratio on the full set as identified in our analysed interactions (36.3 %), there were only 18.8 % of valid questions for which answers were not retrieved.

The total number of requests for answers across all HPD systems was 5,253. Of these, 130 were people typing in the instruction example (i.e., the number 1234), and 598 were known typos, leaving the total potentially valid requests for answers at 4,525. The average number of requests for answers before the answer was ready was 1.16 and (min: 0, max: 18); for after an answer was ready (i.e., repeating the answer) this was 1.44 (min: 1, max: 42). The distribution of times taken to retrieve answers on the HPD system is summarised in Fig. 9. The number of questions asked on one box and answered on another was only 631. While we cannot be certain that some of these are not accidental by typing in random numbers, we would hope that the majority are actual retrieved answers.

Direct question comparisons

As discussed earlier, in order to make direct comparisons with exact audio interactions between systems, we fed 100

random MPI queries into the HPD back-end. Of these, 56 were classed as valid interactions[7], for which the MPI produced 39.3 % relevant responses and the HPD produced 55.4 % relevant responses.

30.4 % of the 56 valid interactions were given non-valid responses by both the MPI and the HPD systems, and 14.1 % given valid responses by both systems. 32.1 % of the questions were given relevant answers by the HPD system but non-relevant responses by the MPI. These particular interactions tended to be questions which were phrased in slightly unusually-constructed sentences, for instance; *"when did Shah Jahan make Taj Mahal for his lover?"* or *"is the western line local train between Bandra to Andheri running or not?"*.

Conversely, nine (16.1 %) questions were given relevant answers by the MPI but not-relevant responses by the HPD. Three of these questions were personal inquiries about the answerer themselves, which were all answered *"I cannot tell you this information"* or *"that is innapropriate"*, while the rest were mostly due to recognition errors on the part of the human answerer. During an additional round of transcription of these interactions, it became clear that the answerers struggled to hear some of the queries, even though they are actually perfectly audible when listened to on desktop speakers. This is potentially an issue that has been extrapolated over the entire data set, and is likely to be a result of the technology used by some of the answerers. That is, crowdsourced answers have two options for responding: a desktop website, or a mobile version. We believe that many of the *"sorry I can't hear you"* responses provided by the HPD answerers were likely a result of low-quality mobile phone speakers.

7 DISCUSSION

Despite installing the same number of each type of system for the same time period, we received over double the number of questions on the MPI systems than on the HPD version. We attribute this response rate to several factors, such the instant nature of the MPI system, which meant that people could stand and ask questions repeatedly and get answers straightaway, plus the lack of relevant responses from the MPI system resulting in the need to re-ask questions.

A further interesting observation is that there are far more non-valid interactions on the HPD system when compared to the MPI (36 % versus 22 %). However, typing (i.e., entering numbers) can be heard in the background on many of these recordings, indicating that these interactions are perhaps attempted answer retrievals as opposed to simple mistakes.

One of the primary differences we saw between the two types of system was the quality and relevance of the answers. We have identified via interaction tagging that the HPD system provides relevant responses to all valid interactions significantly more often than the MPI system, providing 69.1 % and 37.9 % relevant responses, respectively. This trend extends to almost all sub-categories, subjects and interactions. The reason for the low percentage of relevant responses on the machine-powered system is twofold. Firstly, basic *recognition* is poor on the MPI system, which results in many misheard interactions and hence an increased number of irrelevant responses (MPI: 8.4 % verses HPD: 2.1 %), but more often an *"I don't know"* response. We believe the reason for this is not simply down to poor AI, but as a result of the context of use; in particular the public background noise and the range of languages used in queries. As became evident very early on in our tagging process, Dharavi residents, diverse in culture, religion and birthplace, have adapted their conversational speech to include a mixture of languages including Hindi, Marathi, Bengali, Urdu, Tamil and English, which, as previous research has shown [11, 16], current AI systems are struggling to deal with.

Both MPI & HPD speech systems, as we anticipated, had difficulties dealing with noisy environments and multiple people speaking concurrently. Consider the following interaction on the HPD involving two children and one adult speaking (at times) concurrently with motorbikes honking in the background. *Q: "When was [inaudible] born?" A: "Whose birthday do you want to know about? We could not hear you clearly"*. We fed this question audio to the MPI system, which was not able to recognise or transcribe the audio, and received a canned *"I don't know"* response. What is of particular interest here is the different ways in which HPD and MPI systems account for difficulties. We are inspired by the human response that repeats back what *was* understood. This type of 'repair organisation' is a common feature of everyday acts of conversation [22]. In effect, the HPD system feeds back what in computer science terms we might refer to as 'system state'. In Dourish's view *"users of computer systems need to form interpretations of the state of the machine in order to determine appropriate courses of action"* [4]. In the case of the above interaction, the course of action made available is to try and ask the question again. However, in the MPI case it is not made clear if the system *doesn't know* or *couldn't recognise* and therefore doesn't afford the same mechanisms for repair. We suggest that future speech systems feed back to users what was recognised so they can better reason about if and how a question needs to be reposed or rephrased.

The second issue with the MPI system is the *interpretation* of queries. Even if the query was correctly parsed from speech to text, the system still needs to find an answer by performing a web search or by using its own AI repository (e.g., the chicken or egg response). As we have seen via numerous examples, the machine powered system is very poor

[7]This number also excludes several interactions that, due to a 90-minute period of system downtime, were not submitted to human-answerers.

at interpreting questions that are not perfectly structured sentences, more than one question in the same sentence, or questions with no easily-searchable answers, (e.g., predictions, subjective quesions, philosophical queries, etc.). The HPD systems, however, performed better in all cases. To further illustrate this phenomenon, consider the following interaction recorded on an HPD installation: *Q: "Who is the prime minister of America?" A: "There is no prime minister of America. They have a president and his name is Donald Trump"*. Feeding the question audio into the MPI-system led to a canned *"I don't know'* response, even though the system correctly transcribed the audio. Contrasting these responses, we see the human ability to move from what was said to what was *meant* [7], where the person answering the question forms what Balentine, a pioneer of speech and IVRs systems, refers to as a 'theory of mind' [1].

Another notable difference between the two systems is the type of queries received. In general, the MPI systems received far less philosophical, subjective, predictive, advice and personal questions than the HPD systems. We believe this to be a factor of either questioners knowing they are speaking to a person and as a result getting rich answers (in the case of the HPD results) or people learning over time that the responses received for these types of questions are mainly low-quality (in the case of the MPI system). In contrast, the MPI system received more non-questions than the HPD system, in particular requests for music, poems, jokes, etc. as well as more questions directed at the machine itself. The MPI system also had a higher proportion of questions that could be considered "easy to answer" by most in the area (e.g., *"what is the capital of [X]"*, *"who is the leader of [Y]"*, etc.), presumably due to learning effects that show that these types of question are more likely to get a relevant response.

As one might expect from questions directed at a machine versus questions directed at a person, there were significantly more queries in the form of keywords (i.e., search terms) on the MPI system than the HPD system (8.73 % versus 2.56 % respectively). What we did not expect from this, however, is that the use of search terms in the HPD system almost doubled the relevant answer rate when compared to queries that were not classed as search terms (80.8 % relevant for search terms versus 43.4 % relevant for complete sentences). This trend was seen somewhat on the MPI system (i.e., questioners were more likely to get a relevant response using a search term than by using a full sentence) but to a far lesser extent (35.5 % relevant for search terms versus 28.8 % relevant for complete sentences).

8 CONCLUSIONS AND FUTURE WORK

In this paper, we have presented a probe of the use of conversational speech systems in public slum environments. While prior work presented some evidence of the value of such systems in these contexts, it used Wizard-of-Oz approaches for short periods of time. In contrast, the longitudinal evaluation of our two types of systems resulted in a corpus of over 12,000 queries ranging from basic facts to philosophical questions, all the way to practical advice about health, money and well-being. In addition, by deploying both HPD and MPI systems we were able to consider circumstances that might justify the use of a HPD system; or, the development of MPI systems that support types of interaction afforded by the HPD but not yet accommodated in commercial AI-only based approaches.

The engagement of Dharavi residents with the deployed systems provides evidence that public speech installations have value in emerging market contexts, returning accessible information without requiring textual or technological literacy or potentially costly in-home devices. Further, the differences in success rates, types of questions asked and ways of answering seen in the HPD versus the MPI provide useful pointers to developers who wish to elaborate on current state-of-the-art AI approaches.

The current systems leave plenty of scope for refinement, however. For instance, analysis of the MPI data logs show that only 38 % of valid questions resulted in a relevant response, and while the HPD systems performed far better in this area (69 % relevant responses), the nature of the interaction means that there is always a delay in answers being received, making it hard to scale in its current form.

To tackle these issues, we propose a hybrid interface combining the automation benefits of artificial intelligence with the richness of human responses when required. We envisage a triage model where the machine performs a first pass, providing instant answers where possible. If the machine is unable to answer, or if the response is unsatisfactory, the question will then get passed to human respondent. Further, we also propose a machine-learning approach to categorise and reuse human-curated responses, building up a corpus of Q&A pairs over time, and thereby reducing the percentage of frequently wrong answers that plague current AI systems.

Another key area of future work is to move away from remote, crowd-sourced answerers and instead recruit local community members to be the primary human respondents. Our aim is that this shift will not only support flexible employment for question answerers, but may also be able to provide richer, more locally relevant answers in the process. We envisage various ways to ensure answer quality, ranging from expertise rating systems to monetary incentives based on quality. Overall, then, we see the proposed mixed intelligence model as capable of potentially providing a rich trove of high-quality and worthwhile question responses.

ACKNOWLEDGMENTS

We would like to thank Prabodh Sakhardande, Rohit Gupta and Udayan Vidyanta for their help with the deployment studies discussed here. This work was supported by Engineering and Physical Sciences Research Council grants EP/M00421X/1 (https://www.reshapingthefuture.org/) and EP/M022722/1 (https://www.cherish-de.uk/).

REFERENCES

[1] Bruce Balentine. 2007. *It's Better to Be a Good Machine Than a Bad Person: Speech Recognition and Other Exotic User Interfaces in the Twilight of the Jetsonian Age.* ICMI Press, Annapolis, MD, USA.

[2] Frank Bentley, Chris Luvogt, Max Silverman, Rushani Wirasinghe, Brooke White, and Danielle Lottridge. 2018. Understanding the Long-Term Use of Smart Speaker Assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3, Article 91 (Sept. 2018), 24 pages. https://doi.org/10.1145/3264901

[3] Devanuj and Anirudha Joshi. 2013. Technology Adoption by 'Emergent' Users: The User-usage Model. In *Proceedings of the 11th Asia Pacific Conference on Computer Human Interaction (APCHI '13).* ACM, New York, NY, USA, 28–38. https://doi.org/10.1145/2525194.2525209

[4] Paul Dourish. 2004. What We Talk about When We Talk about Context. *Personal and Ubiquitous Computing* 8, 1 (Feb. 2004), 19–30. https://doi.org/10.1007/s00779-003-0253-8

[5] Nick Fox. 2018. The Google Assistant is going global. (2018). Retrieved 3rd September 2018 from https://www.blog.google/products/assistant/google-assistant-going-global/

[6] Google. 2018. Neighbourly: Ask Local Questions & Get Answers. (2018). Retrieved 3rd September 2018 from https://play.google.com/store/apps/details?id=com.google.android.apps.nbu.society

[7] Richard Harper. 2010. *Texture: Human Expression in the Age of Communications Overload.* MIT Press, Cambridge, MA.

[8] Lucy Hattersley. 2018. AIY Voice Essentials. (2018). Retrieved 5th March 2018 from https://www.raspberrypi.org/magpi/issues/essentials-aiy-v1/

[9] Anirudha Joshi, Girish Dalvi, Manjiri Joshi, Prasad Rashinkar, and Aniket Sarangdhar. 2011. Design and Evaluation of Devanagari Virtual Keyboards for Touch Screen Mobile Phones. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI '11).* ACM, New York, NY, USA, 323–332. https://doi.org/10.1145/2037373.2037422

[10] Arun Kumar, Nitendra Rajput, Dipanjan Chakraborty, Sheetal K. Agarwal, and Amit A. Nanavati. 2007. WWTW: The World Wide Telecom Web. In *Proceedings of the 2007 Workshop on Networked Systems for Developing Regions (NSDR '07).* ACM, New York, NY, USA, Article 7, 6 pages. https://doi.org/10.1145/1326571.1326582

[11] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf Between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16).* ACM, New York, NY, USA, 5286–5297. https://doi.org/10.1145/2858036.2858288

[12] Audrey Mbogho and Michelle Katz. 2010. The Impact of Accents on Automatic Recognition of South African English Speech: A Preliminary Investigation. In *Proceedings of the 2010 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists (SAICSIT '10).* ACM, New York, NY, USA, 187–192. https://doi.org/10.1145/1899503.1899524

[13] NPR and Edison Research. 2018. *The Smart Audio Report, Spring 2018.* Technical Report. National Public Media LLC. https://nationalpublicmedia.com/smart-audio-report/

[14] Peter Pirolli. 2007. *Information Foraging Theory: Adaptive Interaction with Information.* Oxford University Press, Oxford, UK.

[15] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18).* ACM, New York, NY, USA, 640:1–640:12. https://doi.org/10.1145/3173574.3174214

[16] Aung Pyae and Tapani N. Joelsson. 2018. Investigating the Usability and User Experiences of Voice User Interface: A Case of Google Home Smart Speaker. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct (MobileHCI '18).* ACM, New York, NY, USA, 127–131. https://doi.org/10.1145/3236112.3236130

[17] Question Box. 2018. Question Box: Overview. (2018). Retrieved 21st September 2018 from http://www.questionbox.org/overview/

[18] Quora. 2018. About Quora – Quora. (2018). Retrieved 21st September 2018 from https://www.quora.com/about

[19] Agha Ali Raza, Rajat Kulshreshtha, Spandana Gella, Sean Blagsvedt, Maya Chandrasekaran, Bhiksha Raj, and Roni Rosenfeld. 2016. Viral Spread via Entertainment and Voice-Messaging Among Telephone Users in India. In *Proceedings of the Eighth International Conference on Information and Communication Technologies and Development (ICTD '16).* ACM, New York, NY, USA, 1:1–1:10. https://doi.org/10.1145/2909609.2909669

[20] Agha Ali Raza, Farhan Ul Haq, Zain Tariq, Mansoor Pervaiz, Samia Razaq, Umar Saif, and Roni Rosenfeld. 2013. Job Opportunities Through Entertainment: Virally Spread Speech-Based Services for Low-Literate Users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13).* ACM, New York, NY, USA, 2803–2812. https://doi.org/10.1145/2470654.2481389

[21] Simon Robinson, Jennifer Pearson, Shashank Ahire, Rini Ahirwar, Bhakti Bhikne, Nimish Maravi, and Matt Jones. 2018. Revisiting "Hole in the Wall" Computing: Private Smart Speakers and Public Slum Settings. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18).* ACM, New York, NY, USA, Article 498, 11 pages. https://doi.org/10.1145/3173574.3174072

[22] Emanuel A. Schegloff, Gail Jefferson, and Harvey Sacks. 1977. The Preference for Self-Correction in the Organization of Repair in Conversation. *Language* 53, 2 (1977), 361–382. https://doi.org/10.2307/413107

[23] Lucy Suchman. 2002. Located Accountabilities in Technology Production. *Scandinavian Journal of Information Systems* 14, 2 (Sept. 2002), 91–105.

[24] Lucy Suchman. 2002. Practice-Based Design of Information Systems: Notes from the Hyperdeveloped World. *The Information Society* 18, 2 (March 2002), 139–144. https://doi.org/10.1080/01972240290075066

[25] Harold Thimbleby. 2013. Reasons to Question Seven Segment Displays. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13).* ACM, New York, NY, USA, 1431–1440. https://doi.org/10.1145/2470654.2466190

[26] Aditya Vashistha, Edward Cutrell, Gaetano Borriello, and William Thies. 2015. Sangeet Swara: A Community-Moderated Voice Forum in Rural India. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15).* ACM, New York, NY, USA, 417–426. https://doi.org/10.1145/2702123.2702191

[27] Marion Walton and Vera Vukovic. 2003. Cultures, Literacy, and the Web: Dimensions of Information "Scent". *Interactions* 10, 2 (March 2003), 64–71. https://doi.org/10.1145/637848.637864

[28] YouGov. 2018. Smart Speaker Ownership Doubles in Six Months. (April 2018). Retrieved 7th August 2018 from https://yougov.co.uk/news/2018/04/19/smart-speaker-ownership-doubles-six-months/