International Journal of Speech Technology
https://doi.org/10.1007/s10772-021-09806-2

Personalized weather information for low-literate farmers using multimodal dialog systems

Muhammad Qasim[1] · Haris Bin Zia[2] · Awais Athar[3] · Tania Habib[1] · Agha Ali Raza[4]

Received: 11 May 2020 / Accepted: 6 January 2021

Abstract
Speech-based services over simple mobile phones are a viable way of providing information-access to under-served populations (low-literate, low-income, tech-shy, handicapped, linguistic minority, marginalized). Despite the simplicity and flexibility of speech input, telephone-based information services commonly rely on push-button (DTMF) input. This is primarily because high accuracy automatic speech recognition (ASR), that is essential for an end-to-end spoken interaction, is not available for several languages in developing regions. We share findings from an HCI design intervention for a dialog system-based weather information service for farmers in Pakistan. We demonstrate that a high accuracy ASR alone is not sufficient for effective, inclusive speech interfaces. We present the details of the iterative improvement of the existing service that had low task success rate (37.8%) despite being based on a very high accuracy ASR (trained on target language speech data). Based on a deployment spanning 23,997 phone calls from 6893 users over 10 months, we show that as multimodal input, user adaptation and context-specific help were added to supplement the ASR, the task success rate increased to 96.3%. Following this intervention, the service was made the national weather hotline of Pakistan.

Keywords HCI4D · ICT4D · Pakistan · Low-literate farmers · Spoken dialog systems · DTMF

1 Introduction

Information and Communication Technologies (ICTs) can provide people access to information in a fast and efficient manner. ICTs are used as developmental tools to provide services to people and catalyze economic growth. Despite the benefits, there are significant barriers to their effective use in developing countries. These include low literacy, lack of skills to handle technology, lack of resources to afford high-end communication devices, and insufficient internet facilities. These barriers need to be addressed to utilize the full potential of ICTs targeting under-served populations.

✉ Haris Bin Zia
haris.zia@itu.edu.pk

Agha Ali Raza
agha.ali.raza@lums.edu.pk

[1] University of Engineering and Technology, Lahore, Pakistan

[2] Information Technology University, Lahore, Pakistan

[3] European Bioinformatics Institute, Cambridgeshire, UK

[4] Lahore University of Management Sciences, Lahore, Pakistan

Automated mobile phone-based, speech services (also called *Interactive Voice Response* (IVR) services) are one of the key techniques to overcome accessibility challenges (Thies 2015). In IVR services, users are provided information access by allowing them to interact with automated systems over regular phone calls. IVR services do not require the target users to have internet access, smart phones, or the ability to read or write. Users are provided their required information via spoken interactions using only simple (non-smart) phones over regular phone calls. These services have seen a rise in information dissemination campaigns targeting low-literate, low-income and tech-shy people (Sherwani et al. 2007; Patel et al. 2010; Mudliar et al. 2012; Raza et al. 2013, 2018, 2019; Vashistha et al. 2015, 2017, 2018; Wolfe et al. 2015; Moitra et al. 2016; Ahmad et al. 2017; Swaminathan et al. 2019; Vashistha et al. 2019a, b). IVR services offer the benefits of requiring no more than simple or feature phones and users' ability to make and receive phone calls and converse in their local language. As a result, these services often outperform SMS and smartphone-based interventions in terms of the scale of spread among low-literate and poorly connected users (Vashistha et al. 2019a, b).

Published online: 01 February 2021

IVR services could be categorized in terms of the interface as *push-button-input, spoken-output* services where the users provide input by pressing keys on the numeric keypad and system speaks out the needed information; and *spoken-input*, *spoken-output* services (aka *Spoken Dialog Systems*) where users are required to provide voice input. While we have seen a rise in the use of the push-button input systems, Spoken Dialog Systems have not been utilized much to design information services in developing countries. This is despite the fact that spoken dialog systems are simpler, more flexible and assume even lesser technical capabilities from users as compared to push button systems. This is due to two reasons: (1) there is a lack of high accuracy speech recognition capabilities for languages of developing regions and (2) there are hardly any interface design guidelines available for Spoken Dialog System for hard-to-reach populations.

High accuracy *Automatic Speech Recognition* (ASR) capability is considered an essential component of Spoken Dialog Systems. As there is a lack of high accuracy ASR capability for languages of developing regions, speech services targeting users in these regions mostly resort to push-button (DTMF) input (Sharma Grover et al. 2009). This is true even in cases where the spoken input is easier to elicit as compared to key-presses. Examples of such contexts include open-ended input questions as well as questions with more than 10 answer choices that cannot be entered easily using a numeric keypad. For example, it is harder to use the push-button input to elicit information about location and profession etc.

The second reason is that there are hardly any guidelines available for designing Spoken Dialog Systems for low-literate users. We know that it is harder to elicit appropriately "formatted" speech input from tech novice users over mobile phones at a high enough signal-to-noise ratio (Sherwani et al. 2007). Speech recognition errors also lead to greater annoyance and confusion among such users (Sherwani 2009; Sherwani et al. 2009). On the other hand, push-button input is more accurate but not as natural as speech input because users constantly need to remove the phone from their ears to press keys (Sherwani 2009).

In this paper, we explore the design and usability of speech interfaces that employ spoken output and multimodal (speech and push-button, input) to provide *inclusive* information access to hard-to-reach populations. Based on lessons from an HCI design intervention into an existing weather service for farmers in Pakistan followed by a 10-month long deployment, we show that high accuracy speech recognition alone is not sufficient for effective speech interfaces for low-literate people. Other modalities and interface features are needed to supplement the ASR to provide a richer experience and higher task success rate. We present a set of recommendations and design principles including multimodal input, user adaptation, and context-specific help.

2 Contributions

There are two main research contributions of this paper: (1) it provides interface design considerations for Spoken Dialog Systems for hard-to-reach users and, (2) it shows that high accuracy speech recognition capability, although important, is not sufficient for effective speech interfaces for hard-to-reach users. The ASR needs to be supplemented with other HCI components to perform better.

We started with an existing dialog system with a reported high accuracy speech recognition back-end that performed very poorly in terms of task success and engagement among users. Our intervention improved this service over several iterations by adding better interface design and user adaptability. The resulting yield is quantified and presented at each step. This paper could also be regarded as a tested checklist of HCI considerations for designing Multimodal Dialog Systems targeting users with special needs (low-literate, visually impaired, non-tech savvy). Concretely, we show that:

- A well-trained and high accuracy speech recognition system with a simple and bug-free interface is not sufficient for providing effective and inclusive information access
- The presented intervention reemphasizes that iterative retraining of speech recognition systems with actual field data leads to significant improvement in speech recognition accuracy (this is true even for systems that already perform at high accuracy)
- Interface design interventions like context specific help with examples, push-button input in case of speech recognition failure and system's ability to adapt to users significantly improve task success rate
- An information service could still perform at a very high task success rate with relatively low-contribution from speech recognition technology

3 Related work

Speech Interfaces allow humans to communicate with computers in their spoken language (McTear 2002). Such systems have been used to provide information access (Bratt et al. 1995; Zue et al. 2000), reservation services (Seneff and Polifroni 2000), maintenance support (Bohus and Rudnicky 2005), tutoring (Litman and Silliman 2004), navigational services (Pellom et al. 2001)

and crowdsourcing (Vashistha et al. 2017). In developing countries where low-literacy is a major hurdle to textual interfaces, speech interfaces have been successfully used in several domains, including health (Sherwani et al. 2007; Wolfe et al. 2015; Batool et al. 2017; Ahmad et al. 2017), disaster response and recovery (Roche et al. 2006), citizen journalism (Ejaz et al. 2018; Gram Vaani 2017; Gulaid and Vashistha 2013; Mudliar et al. 2012), agriculture (Maneesha and Abhishek 2014; Patel et al. 2009, 2010), job search (Raza et al. 2013; White et al. 2012), e-government (Rocheleau and Wu 2005), marketplace (Vashistha et al. 2018; Zainudeen et al. 2010), finance (Rocheleau and Wu 2005), information delivery (Wolfe et al. 2015), entertainment (Raza et al. 2012; Wang and Singhal 2018), social connectivity (Vashistha et al. 2015; Raza et al. 2018) and education (Raza et al. 2019).

Medhi et al. (2006) compare textual and non-textual interfaces for digital maps and job search systems for low-literate users. Conducted in slums of Bangalore, the study highlighted the importance of help options and confirmed that non-textual and voice-based systems are preferred by low-literate users over textual ones.

Speech vs. push-button (DTMF) input is an important question in the design of speech interfaces. Lee and Lai (2005) report a study involving 16 participants who were asked to perform tasks like accessing e-mails, voice mail, and calendar. They conclude that DTMF is more efficient for linear and simple tasks while speech is better for non-linear tasks. Users preferred speech input despite a high word recognition error rate of 20–25% as they found the experience of a machine recognizing their spoken input entertaining and enjoyable.

Grover et al. (2009) report a spoken dialog system for providing health information to caregivers of HIV positive children of Botswana. They compare different input modalities among semi and low literate users. They report task completion rates for speech and DTMF input modes to be similar, and tech-literacy being a more important factor than overall literacy for task completion. They report that 59% of users preferred DTMF while only 19% preferred speech input. Project HealthLine (Sherwani et al. 2007, 2009) found that speech input performed better than DTMF in terms of task completion, for both literate and low literate users.

On the other hand, Patel et al. (2009, 2010) both report that DTMF and numerical input performs better than speech in terms of task completion and performance improvement. They report the problem of transitioning between key presses and speaking as a major challenge of DTMF. Overall, the study suggests that DTMF input is more intuitive and reliable than speech and is a better choice if user perception is vital to system adoption. Zue et al. (2000) present JUPITER, a natural language conversational agent for access to worldwide weather information. JUPITER was not oriented towards or tested on a tech-novice audience.

In terms of speech recognition required for under-resourced languages, Qiao et al. (2010) developed a technique called SALAAM. SALAAM can be quickly trained to perform high accuracy speech recognition for small vocabulary tasks (50–100 words). SALAAM method is used by Cuendet et al. (2013) and Reda et al. (2011) to provide farmers a way to search for agricultural extension videos. In our work, we expected a vocabulary size greater than 100 words (at least 139 to cover all districts in Pakistan). As SALAAM performs best for vocabulary sizes up to 100 words, we did not use it for district name recognition.

4 Original problem description

The work presented in this paper has been done to solve a particular information access problem faced by low-literate farmers in Pakistan: access to localized weather information. The livelihood of farmers is directly linked with weather conditions. Their everyday decisions such as when to sow, irrigate and harvest their crops, administer pesticides etc. are dictated by actual weather conditions. However, most of the farmers being low-literate, there are no simple ways for them to find the weather forecast for their specific locations.

Pakistan is a country of diverse climatic and extreme weather conditions. Various geographical regions of Pakistan experience extreme weather conditions (spells of hot, cold, rainy or dry weather and floods) that keep fluctuating throughout a season. As agriculture is the largest sector of Pakistan's economy i.e. 24% of GDP (*Pakistan Bureau of Statistics* 2019), providing timely and accurate weather information has a direct impact on the economy.

The need to access timely and location-specific weather information is not limited to farmers. Most of the river-irrigated plains annually flood due to heavy rains and melting snow that leads to great loss of life and property. Timely forecast of floods could help people decide when to move to safer areas. Mountain roads get blocked for several months as soon as heavy snowfalls begin and such information could help people leave the highlands in a timely manner. People also use weather forecasts to plan travel and outdoor activities as such activities are severely curtailed by heavy rain, snow and heat waves.

The biggest challenge towards accessing weather information using traditional means is low-literacy. The literacy rate in Pakistan is 57.9% (45.8% among females), which is even lower for rural areas (around 51%) (*The World Factbook-Central Intelligence Agency* 2017). This means that close to half of the population cannot benefit from the textual information. Another major issue is internet availability as broadband subscription stands at 36% of the total population and 3G/ 4G subscribers at 35% which leaves the majority of the population offline (*Pakistan Telecommunication Authority*

2019). On the other hand, mobile penetration in Pakistan has increased drastically over the last decade and currently, there are 161 million active mobile subscribers (80.5% of the population) (*Pakistan Telecommunication Authority* 2019). Hence, providing information related to health, weather, floods, and other emergencies through mobile phones can help in saving lives and improving the living standards of a large number of people.

5 Initial solution

The Pakistan Meteorological Department (PMD) runs a hotline where people can call in and ask for weather information. All calls were answered by human operators who are only available during office hours (9 am–5 pm). PMD was interested in scaling their manual hotline and running it 24/7. A project was conceived by PMD and the original designers of the weather service with the goal of creating an IVR hotline [Weather Information Service (WIS)] that would allow people to call in and access location-specific weather information and forecasts. The biggest technological hurdle in this process was to elicit location information from the callers. One possible solution of using real-time cell tower information was dismissed at a very early stage of the project because of privacy laws.

As it is not practical to elicit the names of hundreds of locations using push button input, they decided to create a spoken dialog system. An automatic speech recognition system was rigorously trained with district names in all popular accents of Urdu: the most widely spoken and understood language in Pakistan. However, once deployed the service performed very poorly in terms of task success and user satisfaction. Call logs and recorded input revealed that only 37.8% of calls on average were able to result in weather information being conveyed to the user.

At this point, a collaboration was formed between the authors of this paper, the designers of WIS and the Pakistan Meteorological Department. An early analysis revealed shortcomings with the user interface of the service and a design intervention was planned. It was decided to improve the interface by adding adaptability, context-specific help, culturally appropriate prompts and multimodal back-off in case the speech recognizer fails. This resulted in a drastic improvement in call success rate to 96.3%. The service was also made the national weather hotline of Pakistan.

6 Initial design and deployment

This section describes the design and deployment of the Weather Information Service (WIS) before intervention and the problems identified with the said interface as part of the intervention. The initial design is referred to as *Stage-1*. The stages would be explained later in the text.

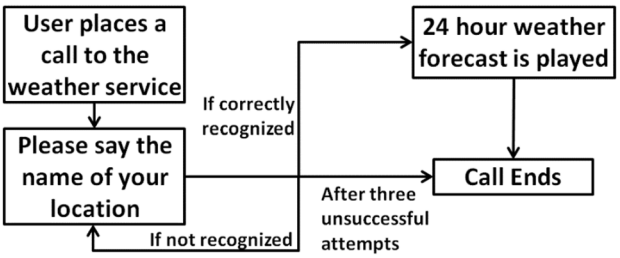

Fig. 1 Initial interface design

6.1 Stage-1: interface design before intervention

The interface of WIS was designed to be simple and usable by tech novice users. It was designed to provide weather information for one location in each call. All interactions were based on incoming calls from users on a normal landline phone number already advertised by PMD and the airtime charges were borne by the users. As shown in Fig. 1, the interaction started when users called the advertised phone number of WIS. Users were asked to speak out the name of the location for which weather information was required. A beep was played after which they had four seconds to record their response. Barge-in was not configured in the dial plan which meant that users could not interrupt the system prompt and had to wait for beep to be played before recording their responses. A voice activity detector isolated out the spoken portion of the recording and the system tried to recognize the recording.

If the input was successfully recognized by the automatic speech recognition (ASR) system, the weather information for the requested location was fetched and played to the users. The information consisted of 24-h weather forecast retrieved from PMD's records. Otherwise, a generic error prompt ("We apologize that the system is unable to understand your input") was played followed by a repetition of the voice prompt that asked the users to speak out the location name. The call automatically disconnected after the third unsuccessful retry or after successfully playing the weather information to the users.

6.2 Deployment

The described interface was fully functional and simple. Speech data consisting of Pakistani district names that was used to train and test the ASR was gathered from all provinces and territories of Pakistan. The districts were selected from official listing of districts as per Pakistan Bureau of Statistics. A district is an administrative unit in Pakistan

which is used to allocate government resources. A district is a part of larger administration unit know as division which in turn is part of a province/territory. Pakistan has a large number of cities and towns ranging from densely to sparsely populated. There are also a large number of towns that are larger in population than some of the cities. There is no clear distinction between what constitutes a city and what constitutes a town. Hence, in order to avoid this confusion, 139 cities that have been officially declared districts have been selected. Details of the IVR system, ASR and the gathered data are described in detail in (Rauf et al. 2015; Qasim et al. 2016a, b). The speech corpus consisted of more than nine hours of speech data (41,443 utterances) recorded from 300 speakers (both male and female) from all over Pakistan covering all six major accents. The data was collected over mobile phones, sampled at 8 kHz and digitized at 16 bits per sample. Hence, the data was aimed to be similar to data expected from the field. In the rest of this paper, we will refer to this speech corpus as DNC (District Name Corpus). The service was deployed on PMD's phone number on August 13, 2015, and received 7683 calls from 2398 users during the first 2 months. Despite the simple interface, the overall call success rate remained very low (only 37.8% successful calls).

We define *call success* as the user's ability to retrieve the required information or being informed of the exact problem (e.g. no input, noisy input, invalid input or input uttered at a very low speaking volume). This implies that the user was able to continue the interaction long enough and WIS was able to successfully decode the input and inform the user accordingly. The call success rate is calculated as:

$$\textit{Call Success Rate} = \textit{Number of successful calls} \,/\, \textit{Potential number of successful calls}$$

The call types that do not count towards the denominator are the ones where the user hangs up before providing any useful input or the ones that fail due to network errors. These are the calls that could not have succeeded even in case of a human operator as the provided input was not meaningful.

At this point, we formed the collaboration and planned our intervention. We analyzed the data gathered from the deployment and investigated the reasons for the low call success rate.

6.3 Reasons identified for poor call success rate

We performed an analysis of all the district name recordings gathered between August 13, 2015, and Oct 05, 2015. The number of recordings per call varies between 0 and 3 depending on the number of attempts it took the system to correctly decode the input (or to give up after 3 attempts). A total of 9966 recordings were gathered from 7683 calls during the initial 2 months.

Next, we present a detailed description of the problems that were identified as the cause of low call success rate.

6.3.1 Poor ASR performance on field recordings

We found that the recognition accuracy of the ASR on actual field recordings came out very low (65%) as compared to its accuracy on the test data of DNC (93%) and the following turned out to be the causes:

6.3.1.1 Multiple words This is where the user speaks out multiple words or a whole phrase instead of (or in addition to) providing the name of the location. The highest percentage of recordings (44.6% out of 9966 recordings) was of this type. Instead of speaking out the name of the district, users uttered complete phrases like: "I would like to know the weather for Karachi", "Lahore (pause) Lahore", "Quetta's weather", "Peshawar, KPK". These phrases were not recognized by the ASR because it had only been trained to recognize isolated words. This indicated that our audience needed help and training for correct usage of the service and for appropriate formatting of their input.

6.3.1.2 Out-of-vocabulary words (OOVs) This is where a user speaks out names of locations that the ASR has not been trained to recognize. At this point, we found a major mistake in the voice instructions of the service. While the ASR had been trained to recognize districts; the actual voice prompt asked the user to speak out the name of a "city". Pakistan has 139 districts (132 districts + 7 tribal agencies) and around 271 cities. The demarcation between what constitutes a city and a village is also not very well defined among the population. As a result, users just spoke out the name of their current location that often did not match any district name.

6.3.1.3 Noisy field recordings As already explained, the ASR although was trained on very *field-like* data (District Name Corpus) and performed very well at recognizing such recordings. However, actual user recordings also contain noise and *disfluencies* (pauses, repetitions, incomplete utterances, filler words like *ummm*, *hmmm*, *aa*, …). The ASR needed to be retrained with user recordings in order to deal with actual field data.

6.3.1.4 Noisy training data A subset of training data was found to be noisily annotated (incorrect spellings, incorrect speech to text mapping etc.). The data needed to be cleaned.

Table 1 Description of stages and HCI interventions

Stages	Length (days)	Details
1	52	Plain ASR with three attempts at recognition (as described in "Interface Design before Intervention")
2	72	Context-specific help was added; Location prompt was fixed; ASR data was cleaned
3	32	ASR was retrained with actual field data; Push-button-based Area Code fallback was added
4	35	Added logging of user locations for adaptability
5	114	Push-button-based Division level fallback was added

6.3.2 Lack of user engagement

The interaction, although precise and simple, was very dry and distant. There were no greetings at the beginning and end of the interaction. The voice prompts were not high quality. They had been recorded on various occasions by two separate female voice artists and as a result, there were abrupt changes in voice and ambient noise.

There was no context-specific help or guidance regarding the nature and cause of recognition errors to alleviate user's anxiety and impatience in case their input repeatedly gets misrecognized. Finally, users were not asked to confirm if they were satisfied with the played weather information. The service just played out the weather as soon as a location recording was successfully decoded, without confirming with the user if the location name was correctly recognized by the system or not.

6.3.3 No fallback in case of speech recognition failures

The service had no fallback options in case user was not able to correctly enter the spoken location information. Ideally, such a service should provide a push-button input or human operator option in case the primary input method fails. This is especially essential when the user is at a place with high background noise or is on a noisy GSM channel or has difficulty pronouncing location name properly.

6.3.4 Lack of adaptability

We found that most users keep requesting weather information for the same location, presumably where they reside. WIS lacked the ability to remember the location last entered by the user and to play its weather information without making the user go through the process of entering a location name again.

6.3.5 Lack of user feedback

There was no way for users to provide us with their feedback and suggestions.

6.3.6 Weather summary

Based on their experience, PMD believed that several callers just call in for a nation-wide weather summary (province-level with special weather events highlighted). These users are often not interested in only the weather specific to their location. Such users were interested in questions like the following: "Whether monsoon (rainy) season has started in the country or not?", "Whether Spring has started in any part of the country", "Has there been recent rainfall in neighboring regions?". WIS did not have a weather summary feature.

Next, we started implementing solutions for each of the issues identified above. These changes were incorporated over a period of 8 months with direct impact on call success rate as described below. Table 1 defines the various "stages" in which new interface features were deployed.

7 Design intervention

This section describes the series of changes that were made to the system and its interface over a period of 8 months to resolve the problems identified in the last section. Figure 2 shows the call success rate of the system after each change as well as the contributions of various interface features to the percentage of successful calls. Figure 3 shows how these changes impacted the distribution of various kinds of recording. It is also notable that new users kept flowing into the system throughout this period (on average 34% of daily users are new) and we are not confounding user training with the impact of interface changes.

In terms of design philosophy, we consider our users as volunteers willing to provide us with their location information. They may even be willing to navigate slightly complicated interfaces as long as they are sure of accessing the weather information that they cannot find otherwise. This direction of thinking allowed us to come up with (1) fallback mechanisms to "salvage" calls where the ASR fails to recognize the location recording, and (2) adaptable interfaces that would ensure that repeat user does not need to go through the toil of making the system recognize the same location input.

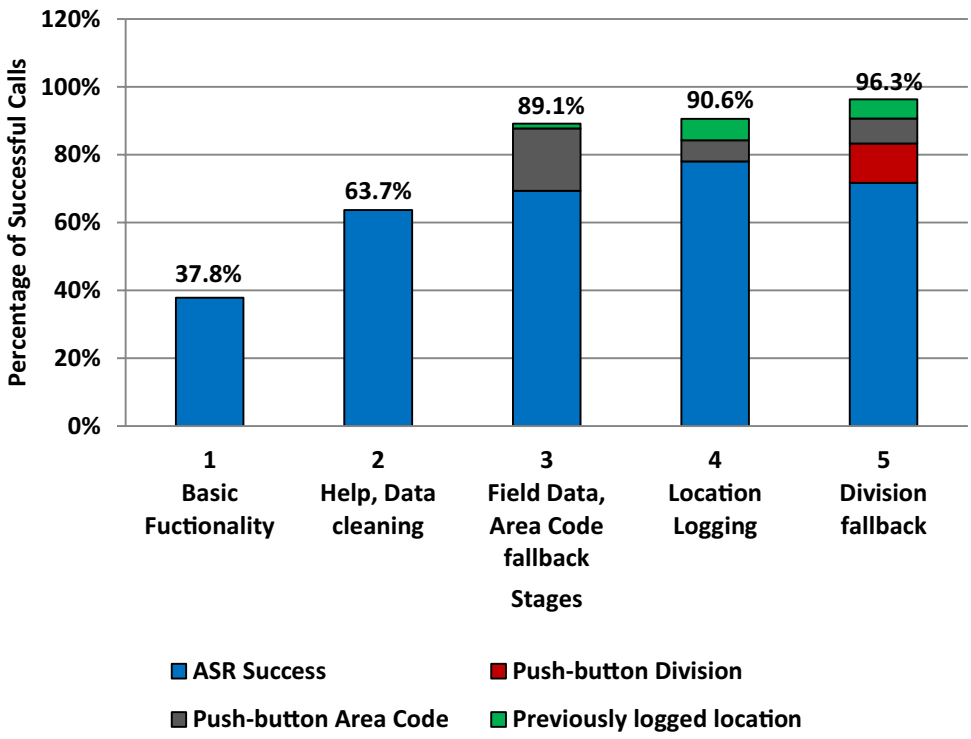

Fig. 2 Call Success rate in various stages

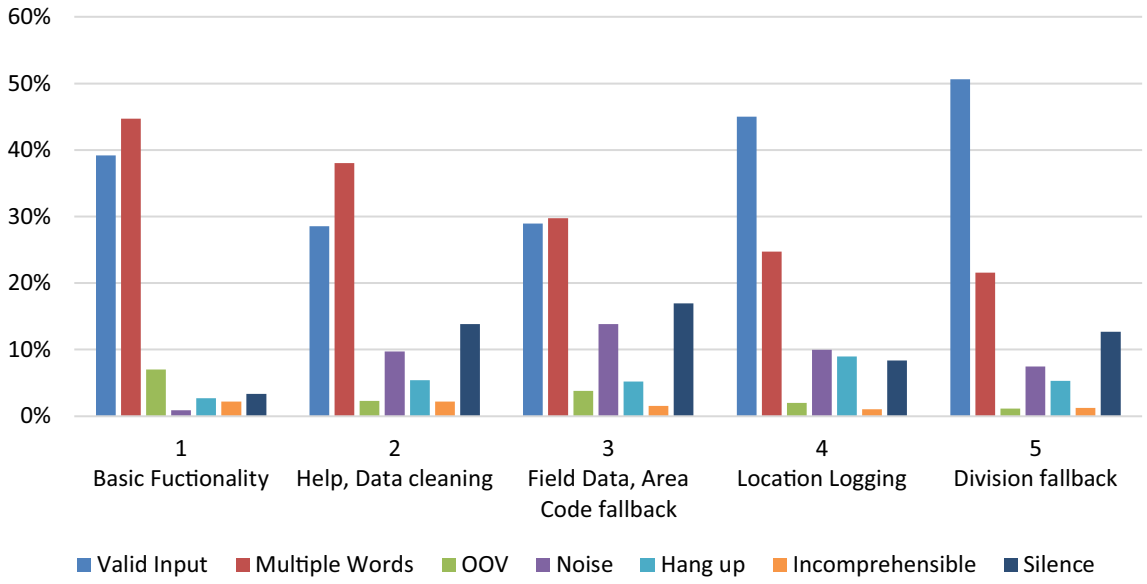

Fig. 3 The distribution of recordings in various stages

7.1 Stage-2: cultural considerations, unambiguous instructions, data cleaning, and context-specific help

The first set of changes was made between Oct 5 and Oct 16, 2015, and included the following. We refer to this as *Stage-2* as shown in Table 1.

7.1.1 Cultural considerations

Traditional welcome greetings (*AsSalaam u Alaikum*, translation: *May peace be upon you*) and farewell greetings (*Khuda Hafiz*, translation: *May God protect you*) were added to the interface. All prompts were rerecorded using high-quality audio by a female voice artist who spoke in a loud, clear voice and at a moderate pace. Urdu was used for

recording all prompts as it is the most widely understood language throughout the country.

7.1.2 Ambiguous voice instructions

The prompt that elicited location information from users was corrected to ask for the name of the *district* of interest instead of a *city*.

7.1.3 Data cleaning

The ASR was retrained after expert annotators scanned all training data and manually removed noisy transcriptions (incorrectly transcribed utterances, spelling errors etc.).

7.1.4 context-specific help with examples

In order to prevent users from uttering multiple words, we trained our ASR to recognize instances of multiple words (a binary decision: does this recording contain multiple words or not?). The accuracy of this recognition improved over several weeks of its initial deployment to 81%. As soon as multiple words are detected, the system plays a prompt informing user that their input has not been recognized because they uttered multiple words and an example correct input is also played as follows:

"We apologize that we are unable to recognize your input. Please *only* say the name of the district for which you would like to know the weather. For example, if you would like to know the weather of Karachi, say *Karachi*!"

We also trained the ASR to automatically detect when a user utters a valid district name multiple times and to recognize it correctly. To prevent users from uttering unsupported location names, we added context specific help to guide users to speak out the correct name. In case of a failure due to these *Out-of-Vocabulary words* (OOVs), the user is informed of the problem and is given an example of correct input:

"We apologize that we are unable to recognize your input. Please say a *valid* district name. For example, if you would like to know the weather of Lahore, please say *Lahore*!"

Similarly, to reduce the number of recordings that get misrecognized because the user does not speak loud enough, we added auto-detection of low recording volume and added a prompt requesting the user to speak louder.

7.1.5 Results

The results of these changes are clearly reflected in Fig. 2 and based on 4,935 calls, call success rate climbed from 37.8% (Confidence Interval (CI) [37.27–38.35%]) to 63.7% (CI [63.04–64.41%]) owing to an increase in the speech recognition accuracy. The accuracy not only improved due to the cleaning of training data but also due to a reduction in the multiple word utterances (from 44.6 to 38%) and OOVs (from 6.9 to 2.2%), calculated based on 10,088 recordings (Fig. 3). The average number of speech recognition attempts per call also increased from 1.36 to 1.83 as the system started doing a better job retaining the users for longer interactions.

Table 2 ASR performance with various types of training data

ASR Trained with District Name Corpus (DNC)		
Training data	33,155 DNC utterances	41,443 DNC utterances
Test data	8288 DNC utterances	586 field utterances
Correctly decoded	7,7093	83
Accuracy	93.01%	65.36%
ASR Trained with a mixture of DNC and Field data		
Training data	33,000 DNC utterances and 6022 field utterances	39,022 DNC utterances and 9,755 field utterances
Test data	7755 DNC utterances and 2000 field utterance	2609 field utterances
Correctly decoded	9133	2415
Accuracy	93.42%	92.56%

7.2 Stage-3: ASR adaptation to field data, push-button fallback, and user feedback

Next, we identified the following updates to resolve more interface problems. We refer to this as *Stage-3* as shown in Table 1.

7.2.1 ASR adaptation to field data

As mentioned earlier the ASR performed very well when tested with recordings from the District Name Corpus (DNC) but poorly on actual field data as shown in the top part of Table 2. An ASR trained with 33,155 DNC utterances and tested with 8288 DNC utterances resulted in an accuracy of 93%. But when trained with 41,443 DNC utterances and tested with 586 field utterances, the resulting accuracy came out to be 65.36%.

We decided to remedy this by transcribing the data gathered from the field and using it to retrain the system in addition to the DNC. The bottom half of Table 2 shows the results. Trained with a mixture of 33,000 DNC utterances and 6022 field utterances and tested with a mixture of 2000 field utterances and 7755 DNC utterances, the system performed at an accuracy of 93.42%. When trained with 48,778 (39,022 DNC + 9755 fields) utterances and tested with 2609

field-only utterances the accuracy still remained at 92.56%. This new ASR was made a part of the weather service.

7.2.2 Push-button fallback to area codes

In order to cater to users who find it hard to appropriately provide speech input to the system (e.g. those who call from a noisy environment or cannot pronounce the names well enough), we deemed important to provide a back off to push-button modality. Although more tedious than spoken input, the push button modality tends to be highly accurate (Sherwani et al. 2007).

The biggest challenge was to find out a way to allow users to specify one of 139 districts using only 10 numerical keys on their keypad. Several options including postal codes, union council numbers, election area codes etc. were considered. Interviews were performed with a sample of target users to find out the input type most popular among them. Finally, landline telephony area codes were chosen as the most commonly known location identifier. We updated the interface to back off to the numerical entry of 3–4 digit landline area codes in case of three failed attempts at spoken input recognition. We also added a second fallback mechanism described later.

7.2.3 User feedback

We added two different user feedback options. A structured feedback prompt after weather information is played to a user:

"If you are satisfied with the weather forecast that has been played to you, press 1; otherwise press 2."

Users who press 2 are directed to the push-button based area code entry option. The second feedback option was added at the very end of interaction where the users are asked to record their feedback and suggestions. This feedback is gathered in the form of unstructured speech and later annotated and reviewed by our annotators. Users are allowed up to a minute to record their suggestions.

7.2.4 Results

As we see in Fig. 2, these changes further improved the call success rate to 89% (N = 1360 calls and CI [88.35–90.03%]). This increase is clearly due to both a major improvement in the spoken input recognition accuracy (69.3% of all calls succeeded because of it) and the back off to area code entry option (responsible for the success of the remaining 18.4% of calls that would have otherwise failed). Figure 3 shows that the multiple words further decreased to 29.7%, however, OOVs increased slightly to 3.7%.

7.3 Stage-4: adapting to the users

An analysis of call data showed a significant trend for users to ask for the same location each time they call. We found that in 37% of 15,648 calls, users had asked for locations that they had inquired about previously. This makes sense as a typical farmer would remain interested in the location of his crops. We learned about the locations-of-residence of 289 users of the service through telephonic surveys (described in section "User Surveys"). We found that these users asked about the weather of their home district in 28.9% of their 280 calls to the weather service. However, WIS interface required users to go through the process of entering their input location each time they called. Therefore, in order to make our system adapt to the locations of our users, we added phone number based user profiles and associated the last successfully recognized location with that phone number. As a result, once users successfully input a location, a summary of the forecast for that location gets played at the beginning of their subsequent calls before they are asked to speak out the name of a district.

Our system currently remembers only the last successfully entered location. This could, however, be easily extended to a model where several locations are associated with a particular phone number (e.g. to cater to users who reside in a location different from the one where they conduct business).

7.3.1 Results

We find that after the addition of this feature call success rate remained stable at 90.6% (N = 2225 calls), however now 78% of calls were successful due to correct recognition by the ASR and a reduced 6.3% were successful due to area code back off. The remaining 6.3% were successful because the users hung up after listening to the weather information of their last known location (Fig. 2). We also see in Fig. 3 that the multiple words further decreased to 24.7% and the OOVs to 1.9%. We also see a decrease in the percentage of incomprehensible recordings to a mere 1.0%. While there is a confounding factor that repeat users may learn on their own to avoid these mistakes, however, we did not find this to be significant. One of the reasons is that on average 34% of all daily users are new. Also, if when we plot these stats (MV, OOVs etc.) within each stage (plots not shown), we do not find any significant reduction across the period of each stage.

7.4 Stage-5: alternate fallback mechanism

We found, based on user feedback and analysis of data (presented in the discussion section) that a large fraction of users were unaware of the landline area codes of their locations. Some remembered obsolete codes. We decided to try alternate ways to fetch actual location names from the users using the keypad.

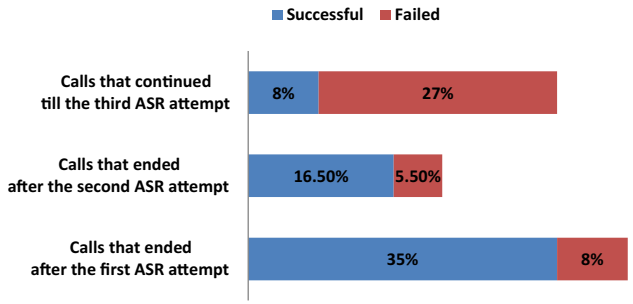

Fig. 4 Call success at various ASR attempts

7.4.1 DTMF divisions

Our biggest challenge was that we could not fit 139 district names on a 10 digit keypad. One possibility is to increase the depth of the input menu and provide the nine most popular district names followed by "for more options press 0". However, this would have required users whose districts were least popular to choose the "more" option 15 times before finally getting to their required district.

We, therefore, decided to go ahead with a different geographical classification. Pakistan has 5 provinces/states and each of the provinces has less than 10 divisions (a division consists of 3–5 districts). Therefore, by compromising on the granularity of the location we were able to gather push-button based location input using just a two levels deep menu. The user was required to first choose the province/state and then the required division.

We also decided to make sure that all unique users consistently back off to only one of these two modalities: areas codes or divisions. These modalities were assigned to the users randomly so that we could meaningfully compare the contributions of these two modes towards call success rate. The randomization algorithm made sure of balanced distribution of users across the two treatments. We assigned each user to the next treatment group (area code or divisions) in a round robin fashion when they reached the error state (their input was not recognized by the ASR multiple times or they indicated that they have not been able to find the weather of their desired location) for the first time. As the system has no control over the order and timing in which the users call it, the round robin alternate assignment adds users to each of the two groups randomly.

7.4.2 Reduced ASR attempts

We also observed that the number of attempts to recognize spoken input provided diminishing returns and also led to a lot of frustration as users had to speak out the name of the district repeatedly. Figure 4 shows that out of the 16,679 calls between stages 1 through 4, 43% (7172) ended after the first ASR attempt, 22% (3669) ended after the second ASR response and the remaining 35% (5838) continued all the way till the third attempt at speech recognition. However, in terms of success, 35% (5838) of all calls (16,679) succeeded in the first ASR attempt, 16.5% (2752) succeeded in the second attempt and only 8% (1334) got successfully recognized in the third ASR attempt (the remaining 4505 calls backed off to area code input).

Clearly the chances of a correct recognition drop significantly as the system reattempts speech recognition. It is also clear that a significant percentage of calls continue till the third ASR attempt (35% of all calls) of which only 22.8% (8% of all calls) is successfully decoded in the third attempt. It was therefore decided to reduce the number of ASR attempts to two before backing off to one of area code or division input to reduce user frustration and salvage most calls.

7.4.3 Results

As shown in Fig. 2, the call success rate in this stage climbed to 96.3% (N = 5731 calls and CI [96.09–96.58%]) from 90.6% (CI [89.99–91.23%]). Now 71% of the calls succeeded due to correct speech recognition, 11.6% succeeded due to division back off, 7.3% succeeded due to area code back off and the remaining 5.6% succeeded as the users hung up after listening to the weather information of their previously inquired district. The multiple words further reduced to 21.5%, OOVs to 1.1% and incomprehensible input remained at a very low 1.2% (Fig. 3).

8 Final interface design

As a result of the above changes, we ended up with the interface shown in Fig. 5. Calls start with traditional greetings followed by a brief country-wide weather summary that highlights unusual weather activity. Return users now hear the weather for their most recently asked district. Next, users are asked to say out the required district name. In case of a recognition failure, they are informed about the exact cause with examples of correct input. Users are allowed up to two attempts of spoken input.

After successful recognition, the weather information is played and users are asked if the system understood their input correctly. Satisfied users are asked for feedback before the calls end. In case of multiple recognition failures or dissatisfied users, the system backs off to push-button input: area codes or divisions.

Users key-in their 3–4 digit landline area code and hear the required weather information. In the case of invalid input, users are allowed up to 2 attempts. If the failure

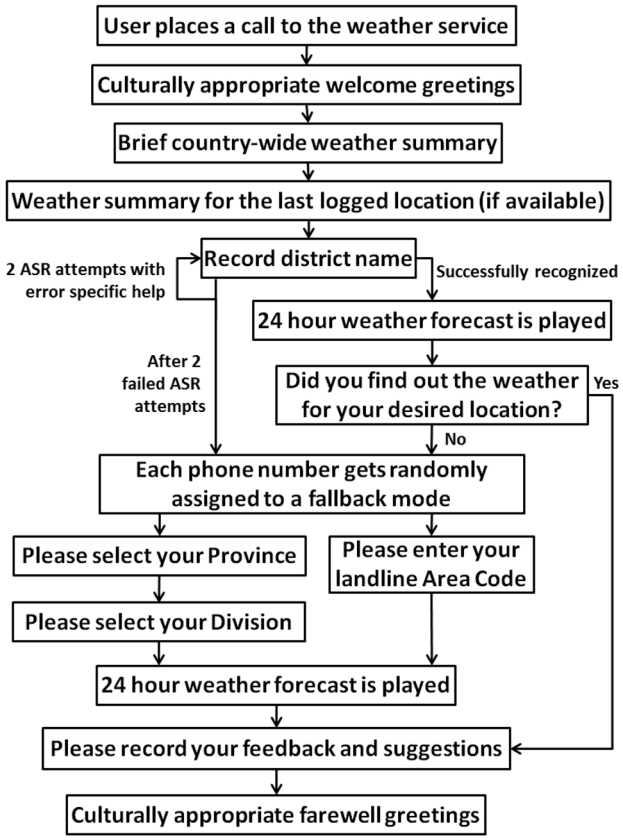

Fig. 5 Final interface design

persists, the system apologizes at not being able to recognize the input and users are asked for feedback. In order to select divisions, users first choose a province (5 choices) and then a division (2 to 9 choices) before hearing the weather information. In case of incorrect keypresses or no input, users are informed about the error and are asked to try again for up to two attempts. In case of a permanent failure, the system apologizes. All calls terminate after asking for feedback, thanking the user and traditional greetings.

9 User surveys

We called up 580 randomly selected users for a telephonic survey. Out of these 335 answered and agreed to participate in the survey. The survey consisted of 21 questions as shown in Table 3. Questions 1–5 and 21 were designed to determine the demographics, literacy, and socioeconomic status of the participants. The question regarding monthly mobile expense was asked to get an estimate of socioeconomic status. Participants were not asked about their gender as this question is considered offensive in the local context (participants get offended why the surveyor is unable to determine their gender based on their voice). Participants' gender was noted down by the surveyor on a binary scale. Questions 6, 12–15 were designed to gauge the technology use and access to alternate sources of information. Questions 7–9 were asked to find out the fraction of users who

Table 3 Telephonic survey questionnaire

#	Question
1	What is your educational background and profession?
2	How old are you?
3	In which district do you reside?
4	Which language do you speak with your family?
5	Several users of the weather service are blind. Are you blind?
6	What type of phone do you use? Touchscreen (smartphone) or one with a keypad (feature phone)?
7	What is the number of your Union Council?
8	What is your PTCL area code?
9	What is the name of your division?
10	How is weather information useful for you? Why do you use the weather service?
11	How frequently do you use the weather service?
12	Do you use the internet?
13	Do you use Facebook or WhatsApp?
14	Other than this service, how can you find out about weather information?
15	Is the weather service better than other sources of information or not? (Participants are asked to specify the reasons for their answers)
16	Does the service recognize the district names you tell it? (Yes/ No/Annoying/Not sure)
17	If it does not recognize your spoken input, what does it ask you?
18	Is the division/area code option better than the speech input? (Yes/No)
19	What do you like about the service? Any specific feature?
20	What do you Like to be improved about the service?
21	What is your approximate monthly mobile expense?

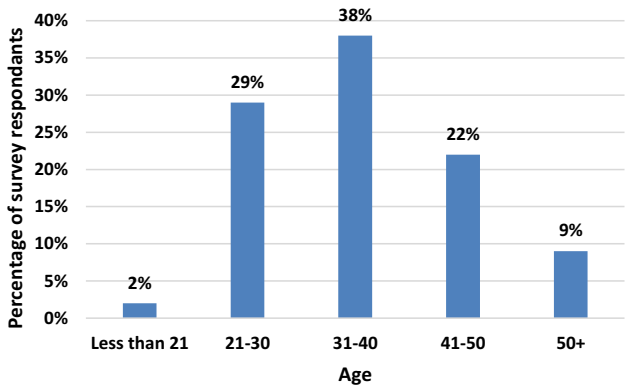

Fig. 6 Age distribution of survey respondents

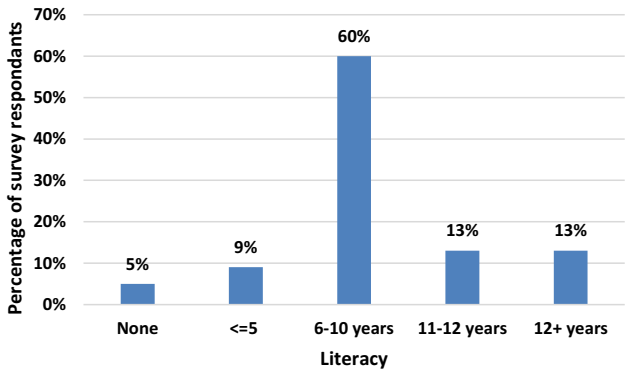

Fig. 7 Literacy distribution of survey respondents

know their union council number, landline area code, and division. Instead of posing these as yes/no questions, participants were asked to spell out this information to determine if they are indeed confident about these details. Finally, questions 10, 11, 16–20 were designed to get usability feedback.

We found that only 3 out of our 335 respondents were females. This came as no surprise as we expected most of our users to be farmers and did not expect a large fraction of females among this group. Their ages were distributed across a wide spectrum as shown in Fig. 6. It was surprising to see 22% of our users with ages 41–50 years and 9% above the age of 50. These ages are not considered typical for new technology adoption. However, this also shows that even very old users learn to use technology given that it provides them with useful information. The 289 users who responded to our question regarding their district, belonged to 49 different districts. In terms of literacy, (Fig. 7), most of our users had 6–10 years of education, which is typical for farmers in Pakistan. There were 14% users with less than 5 years of education and 26% with more than 11 years of education. These people were mostly either landowners or people who were not associated with agriculture. This provides evidence for the usefulness of such services for people beyond the primary intended audience. 71% of our respondents had never used internet, and 75% had never used WhatsApp and Facebook. Television was reported by 52% of them as their alternate source of weather information. However, television neither provides information that is neither highly detailed nor customized to their needs and locality. 25% of the participants accessed weather information over smartphones, while the rest reported word-of-mouth, radio, and newspapers as their alternate source of weather information. 77% said that the weather service better serves their needs compared to other sources of information. When asked for reasons, 84% praised its accuracy compared to other sources, 14% pointed out that it allows them to access weather information whenever needed compared to TV, radio and newspapers which have fixed schedules. The rest pointed out specific features like weather summary.

73% of all survey participants possessed simple and feature phones, while the rest had smartphones (locally referred to as touchscreen phones). There were 67% farmers among the survey participants. This confirmed our hypothesis that personalized weather information is most useful for people directly related to agriculture. Of the remaining 33%, we found people associated with manual labor and construction (7%), shop owners, businessmen, teachers, students and drivers (14.4%) and a handful of carpenters, mechanics, and porters. 2.4% of the users were unemployed. 71% of all contacted users, reported a monthly mobile expense of 500 Pakistani rupees (approximately $4), while 90% reported less than 1000 rupees ($8).

The 290 participants who answered our question about their district belonged to 49 different districts. 72% of these districts were from the Punjab province, followed by 15% from Khyber Pakhtunkhwa, 8% from Sindh and the rest from other provinces. In terms of linguistic diversity, we found that 56% of the contacted users conversed with their families in Punjabi, followed by 23% Saraiki, 6.3% Pashto, 5.65% Urdu, 4% Hindko and rest in Sindhi and Balochi. As Punjabi and Saraiki are predominantly spoken in the Punjab province while Pashto is spoken in Khyber Pakhtunkhwa, these results confirm that most of our users come from the agricultural districts in Punjab and the northern areas of Pakistan. It is very interesting to note that only 5.65% of these users reported Urdu as being their language of everyday conversation. This clearly shows that Punjabi and Saraiki are better options compared to Urdu as interface languages for the weather service.

With regards to their knowledge of location identifiers, we found that 61% of the respondents could tell their Union Council number, 94% knew the name of their division and 60% knew their landline area code. Therefore, division is a good fallback when district name is not recognized. There is no clear winner between union council number and landline

area code. However, there is a caveat that we could not verify whether the self-reported identifiers were indeed correct as we did not have any way of verifying the location of survey respondent (beyond their self-reported locations).

In terms of service usage, 62% of all respondents mentioned that they use the weather service for professional reasons, while the rest reported using it for general reasons. 21% of the respondents said that they use the service at least once daily, 42% reported using it once or twice a week, while 37% only on need basis.

Usability feedback was generally good. Except for 2 respondents, all other (268) respondents said that the service is generally able to recognize their voice input. Only 4% of all respondents were happy with the DTMF fallback while everyone else seemed more satisfied with the speech input. The reason they mentioned was that the spoken input is more convenient. These results are well aligned with Sherwani et al. (2009) who found that DTMF although leads to better input accuracy as compared to speech input, yet fails to yield higher user satisfaction. 97% of respondents considered weather summary and the ability of the weather service to remember their previous location, useful features. Most of the respondents praised the service however, several users also complained about the actual weather information being inaccurate at times. Users also suggested new features like weather forecasts spanning more than a week, weather advisory over SMS for people who are out of cell coverage during parts of the day, and wind speed forecasts.

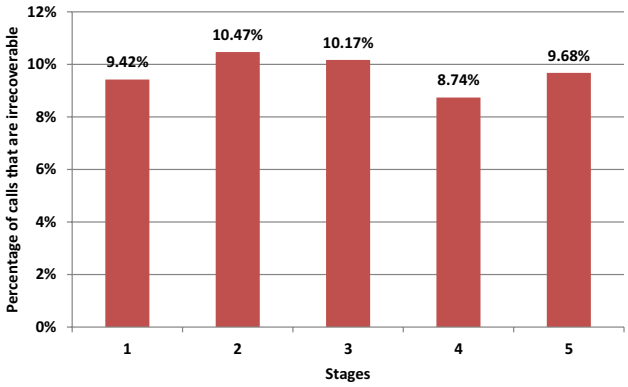

Fig. 8 The percentage of irrecoverable calls

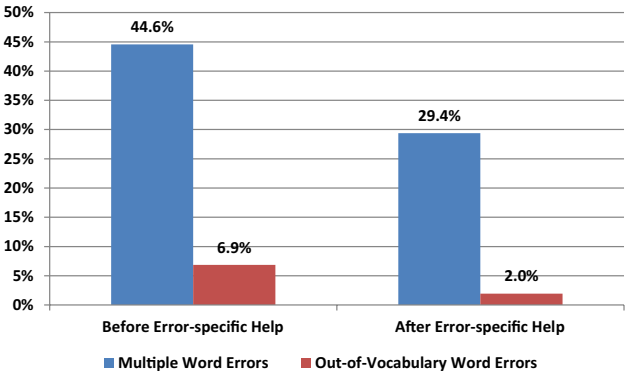

Fig. 9 Reduction of OOV and MW errors

10 Discussion and lessons

In this section, we will isolate the impact of interface changes and compare across interventions.

10.1 Irrecoverable call errors

Between 8.7 and 10.4% of all incoming calls have irrecoverable errors such as calls in which users do not provide any input, there are network or call connection errors, users hang up too soon, or recordings that just contain noise, are all considered to be irrecoverable. Figure 8 shows the percentage of these calls over the stages of deployment.

10.2 Help for multiple word errors and OOVs

Figure 3 shows a gradual reduction in multiple word errors and OOV words over the stages. The gradual improvement is because the multiple word and OOV detection accuracy improved each time we retrained the ASR with more field data. Figure 9 shows the overall comparison of the percentage of recordings with these errors before (out of 10,131 attempts) and after (out of 25,181 attempts) the error specific help and example prompt was introduced.

10.3 Comparison of various forms of push-button fallback

The two different forms of push-button fallback mechanisms (area codes and divisions) were introduced to recover from speech-recognition errors. We made sure that each phone number was randomly assigned to one of these two modes and this assignment was never changed to draw comparisons.

The area codes option remained active for 6 months. Overall 2172 calls backed off to this option after ASR failures. Of these, 884 (40.6%) were correctly recognized through area code. The back off to division option remained active for 4 months. 862 calls backed off to divisions input and of these, 780 (90.4%) were correctly recognized.

We investigated the reasons for the relatively low success rate of area code input and found it to be lack of standardization. Several locations have distinct area codes even though they are not districts. We conducted a study to map all valid area codes entered by users to their respective districts. We

also found that several of our users did not know their area codes.

Both fallback mechanisms have their own pros and cons: area code fallback preserves the granularity of information and does not require users to listen to detailed menus. However, it requires them to know their area codes and also has a prerequisite that multiple districts should not share common area codes.

Back off to divisions is very accurate and does not require users to know anything other than their own division. However, it reduces the granularity of location as we cannot meaningfully fit 139 districts on the numeric options of the keypad and have to resort to a higher level administrative unit. We can of course, increase the depth of the menu further and ask for the district information (3–5 per division). However, this was avoided as it increases user interaction time resulting in an increase in the cost of calls and also makes the interaction more complicated.

10.4 Logging user locations

This feature proved very useful in quickly conveying the needed information to a large fraction of users. In 3040 calls, the weather for a previously recognized district was played out of which, in 2853 (93.8%) calls users hung up immediately afterward, without providing any further input. This makes a lot of sense for regular users who just want to know the updated forecasts for their locations.

In the future, we also plan to experiment with making this option more sophisticated and remembering more than one previous location for each caller. Users can choose from a list of these locations via keypresses (or speech input) at the beginning of each call. However, such levels of customization with the target audience could also prove too complicated.

10.5 Beep vs. no beep

Dialog Systems often prompt users to begin speaking "after the beep". We wanted to understand whether this has any impact on the quality and completeness of speech input for our target audience. We performed a randomized controlled trial where a randomly selected set of 379 users was always instructed to speak "after the beep", while another set of 379 users was always asked to "speak", without any mention of a beep. Barge-in was not configured in both cases. The trial continued for 2 weeks involving 1562 phone calls.

The ASR was successful in decoding 55.8% (CI[54.01–57.59%]) of recordings from users who were never played a beep before recording while it correctly recognized 60.25% (CI [58.51–61.99%]) of the recordings for users who were always asked to record after the beep. We considered ASR success as it depends on the quality and completeness of recordings. Although there is no major difference between the two arms, we still decided to use "speak after the beep" version in WIS as it is most commonly used in Dialog Systems.

10.6 No airtime subsidy

Most of the speech services for low-literate users are designed using a "missed call" mechanism or toll-free numbers to subsidize call airtime costs (Agha Ali Raza et al. 2013; Vashistha et al. 2015). It is also reported that the usage drops significantly as soon as these services discontinue airtime subsidy. In our case, however, the service never subsidized airtime and users paid for the cost of calls at standard rates (approximately 2 cents a minute). Despite that, the usage remained fairly high. This shows that when there are utility and possible financial benefits, users are ready to bear airtime costs. Another reason could be that users did not expect the service to be subsidized as it was never introduced as a free service in the first place.

10.7 High accuracy speech recognition

During stage 1, the ASR performed at an accuracy of 93% on DNC test data. Despite this reported high accuracy, the call success rate came out a mere 37%. Even after data cleaning and retraining with field data in stages 2 and 3 when the recognition accuracy of the ASR increased to 92% for actual field data, it still remained responsible only for the success of 64% to 69% of calls. The rest of the successful calls succeeded due to push-button fallback mechanisms. This clearly shows that even a highly accurate ASR has a limited contribution to an overall successful Dialog System and the user interface must also be inclusive to the target users and their needs.

It is also interesting to note that the push-button division's option was able to successfully provide users with correct weather information in 90.4% of the 862 calls that reached that option. So, even with a poor ASR, this option alone could have led to a high task success rate. We also find that in stage 5, when the call success had increased to 96.3%, only 71.7% of the calls succeeded due to the success of speech recognition. The rest were successful because of push-button fallback mechanisms and logged location information. A very high accuracy ASR, although very important, is mostly not a sufficient ingredient of an effective Dialog System for low-literate users. This is certainly true in case of simple tasks where isolated-word speech recognition could be replaced by cleverly restricting the input to the numeric keypad.

10.8 Choice of interface language

Our telephonic user surveys revealed that Urdu was not the language of everyday communication for 94.35% of our survey respondents. Even though Urdu is the most widely understood language in Pakistan (a country of 71 different languages), we plan to incorporate multilingual interfaces in our future deployments. Next to Urdu, the mostly widely spoken language in Pakistan include Punjabi, Saraiki (in Punjab), Sindhi, Baluchi, and Pashto. Most of the district names remain same or similar across these languages however, the pronunciations of these names depict a lot of variation. Such variations may also account for the low initial and overall ASR accuracy on field data. Our models might have tuned better if trained for a single language (and accompanying district name pronunciations). As a first step, we would need to enrich our dataset with such pronunciation variations. One mechanism of enabling multilingual interfaces is to present a menu of language options at the start of the interaction for all new users, and then use a recognition model based on user's choice. Another method is automatically recognizing the native language (L1) of the user based on their Urdu (L2) recordings and choose the appropriate district-name recognition model.

10.9 User feedback

At the end of each call, the system asks each user to record feedback and suggestions. Of the 2288 feedback recordings, most (91%) were empty (noise, silence, incomprehensible). Of the remaining 200 recordings, 64% contained praise and expression of satisfaction. Some users also expressed why the service is useful for them (it allows them to plan irrigation, application of pesticides etc.). 5% expressed dissatisfaction mostly because of weather conditions not being as forecasted. In another 15% of recordings, users just spoke out names of locations for which they wanted to know the weather.

Most of the remaining recordings (16%) contained very useful feedback and suggestions e.g. to include a country-level weather report (that was later added); to present weather forecast for several days (instead of just 24 h); to provide more detailed weather information and include chances of rain; and to also provide weather information at sub-district level granularity.

11 Conclusion and future work

In this paper, we report an HCI design intervention to improve the task success rate of an existing weather information service for low-literate farmers in Pakistan. Despite very high accuracy speech recognition capabilities, the pre-existing weather service performed poorly at task success. Our intervention identified and improved interface shortcomings that had prevented the service from becoming more simple, robust and inclusive to tech-novice users. We identified several features including multimodal input in case of speech recognition failures, adaptability to ensure that the service caters to the needs of repeat users, and context-specific help as the crucial factors impacting task success. We show that high accuracy speech recognition capability for small vocabulary systems, although important, is not sufficient for effective speech interfaces for hard-to-reach users. In such cases, the speech recognizer can be supported with a push-button fall back to salvage calls that would have otherwise failed. Following our intervention, the service has been the national weather hotline of Pakistan for the last two years.

In the future, we plan to incorporate the SALAAM method to elicit the division-level location information using spoken input instead of DTMF. We also plan to find out the impact of adding a third level to the hierarchy where the user chooses a district within the selected division. Another feature that we want to add to the interface is to remember several locations previously asked by users and to give them the option to select from a list of these locations using speech or DTMF input.

Data availability Speech corpus is available publicly[1].

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Ahmad, S. S. O., Naseem, M., & Raza, A. A. (2017). Maternal awareness for low-literate expecting parents via voice-based telephone services. HCI.

Batool, A., Razaq, S., Javaid, M., Fatima, B., & Toyama, K. (2017). Maternal complications: nuances in mobile interventions for maternal health in Urban Pakistan. In *Proceedings of the Ninth International Conference on Information and Communication Technologies and Development* (p. 3). ACM.

Bohus, D., & Rudnicky, A. (2005). LARRI: A language-based maintenance and repair assistant. *Spoken Multimodal Human-Computer Dialogue in Mobile Environments*, 203–218.

Bratt, H., Dowding, J., & Hunicke-Smith, K. (1995). The SRI telephone ATIS system. In *Proceedings of the Spoken Language Systerns Technology Workshop* (pp. 218–220).

Cuendet, S., Medhi, I., Bali, K., & Cutrell, E. (2013). VideoKheti: Making video content accessible to low-literate and novice users.

[1] http://www.cle.org.pk/clestore/speechcorpus.htm

In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2833–2842). ACM.

Ejaz, H., Hussain, S. A., & Raza, A. A. (2018). The case for IVR-based citizen journalism in Pakistan. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct* (pp. 87–94). ACM.

Gram Vaani. (2017). Retrieved from http://www.gramvaani.org/.

Grover, A. S., Plauché, M., Barnard, E., & Kuun, C. (2009). HIV health information access using spoken dialogue systems: Touchtone vs. speech. In *Information and Communication Technologies and Development (ICTD), 2009 International Conference on* (pp. 95–107). IEEE.

Gulaid, M., & Vashistha, A. (2013). Ila Dhageyso: an interactive voice forum to foster transparent governance in Somaliland. In *Proceedings of the Sixth International Conference on Information and Communications Technologies and Development: Notes-Volume 2* (pp. 41–44). ACM.

Lee, K. M., & Lai, J. (2005). Speech versus touch: A comparative study of the use of speech and DTMF keypad for navigation. *International Journal of Human-Computer Interaction, 19*(3), 343–360.

Litman, D. J., & Silliman, S. (2004). ITSPOKE: An intelligent tutoring spoken dialogue system. In *Demonstration papers at HLT-NAACL 2004* (pp. 5–8). Association for Computational Linguistics.

Maneesha, V., & Abhishek, B. (2014). Innovative IVR system for farmers: Enhancing ICT adoption.

McTear, M. F. (2002). Spoken dialogue technology: Enabling the conversational user interface. *ACM Computing Surveys (CSUR), 34*(1), 90–169.

Medhi, I., Sagar, A., & Toyama, K. (2006). Text-free user interfaces for illiterate and semi-literate users. In *Information and Communication Technologies and Development, 2006. ICTD'06. International Conference on* (pp. 72–82). IEEE.

Moitra, A., Das, V., Vaani, G., Kumar, A., & Seth, A. (2016). Design lessons from creating a mobile-based community media platform in Rural India. In *Proceedings of the Eighth International Conference on Information and Communication Technologies and Development* (pp. 1–11).

Mudliar, P., Donner, J., & Thies, W. (2012). Emergent practices around CGNet Swara, voice forum for citizen journalism in rural India. In *Proceedings of the Fifth International Conference on Information and Communication Technologies and Development* (pp. 159–168). ACM.

Pakistan Bureau of Statistics. (2019). Retrieved from http://www.pbs.gov.pk/content/agriculture-statistics.

Pakistan Telecommunication Authority. (2019). Retrieved from https://www.pta.gov.pk//en/telecom-indicators.

Patel, N., Agarwal, S., Rajput, N., Nanavati, A., Dave, P., & Parikh, T. S. (2009). A comparative study of speech and dialed input voice interfaces in rural India. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 51–54). ACM.

Patel, N., Chittamuru, D., Jain, A., Dave, P., & Parikh, T. S. (2010). Avaaj otalo: a field study of an interactive voice forum for small farmers in rural india. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 733–742). ACM.

Pellom, B., Ward, W., Hansen, J., Cole, R., Hacioglu, K., Zhang, J., et al (2001). University of Colorado dialog systems for travel and navigation. In *Proceedings of the first international conference on Human language technology research* (pp. 1–6). Association for Computational Linguistics.

Qasim, M., Hussain, S., Habib, T., & Rahman, S. U. (2016a). Spoken dialog system framework supporting multiple concurrent sessions. In *2016 Conference of the Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (OCOCOSDA)* (pp. 116–121). IEEE.

Qasim, M., Nawaz, S., Hussain, S., & Habib, T. (2016b). Urdu speech recognition system for district names of Pakistan: Development, challenges and solutions. In *2016 Conference of the Criental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA)* (pp. 28–32). IEEE.

Qiao, F., Sherwani, J., & Rosenfeld, R. (2010). Small-vocabulary speech recognition for resource-scarce languages. In *Proceedings of the First ACM Symposium on Computing for Development* (p. 3). ACM.

Rauf, S., Hameed, A., Habib, T., & Hussain, S. (2015). District names speech corpus for pakistani languages. In *2015 International Conference Oriental COCOSDA Held Jointly with 2015 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)* (pp. 207–211). IEEE.

Raza, A. A., Milo, C., Alster, G., Sherwani, J., Pervaiz, M., Razaq, S., et al. (2012). Viral Entertainment as a vehicle for disseminating speech-based services to low-literate users. In *International Conference on Information and Communication Technologies and Development (ICTD)* (Vol. 2).

Raza, A. A., Saleem, B., Randhawa, S., Tariq, Z., Athar, A., Saif, U., et al. (2018). Baang: a viral speech-based social platform for under-connected populations. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (p. 643). ACM.

Raza, A. A., Tariq, Z., Randhawa, S., Saleem, B., Athar, A., Saif, U., et al. (2019). Voice-based quizzes for measuring knowledge retention in under-connected populations. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (p. 412). ACM.

Raza, A. A., Ul Haq, F., Tariq, Z., Pervaiz, M., Razaq, S., Saif, U., et al. (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* (pp. 2803–2812). ACM.

Reda, A., Panjwani, S., & Cutrell, E. (2011). Hyke: a low-cost remote attendance tracking system for developing regions. In *Proceedings of the 5th ACM workshop on Networked systems for developing regions* (pp. 15–20). ACM.

Roche, R., Hladilek, E., & Reid, S. (2006). *Disaster recovery virtual roll call and recovery management system*. Google Patents.

Rocheleau, B., & Wu, L. (2005). E-Government and financial transactions: Potential versus reality. *The Electronic Journal of E-Government, 3*(4), 219–230.

Seneff, S., & Polifroni, J. (2000). Dialogue management in the Mercury flight reservation system. In *Proceedings of the 2000 ANLP/NAACL Workshop on Conversational systems-Volume 3* (pp. 11–16). Association for Computational Linguistics.

Sharma Grover, A., Stewart, O., & Lubensky, D. (2009). Designing interactive voice response (IVR) interfaces: Localisation for low literacy users.

Sherwani, J. (2009). *Speech interfaces for information access by low literate users* (PhD Thesis). Carnegie Mellon University.

Sherwani, J., Ali, N., Mirza, S., Fatma, A., Memon, Y., Karim, M., et al. (2007). Healthline: Speech-based access to health information by low-literate users. In *Information and Communication Technologies and Development, 2007. ICTD 2007. International Conference on* (pp. 1–9). IEEE.

Sherwani, J., Palijo, S., Mirza, S., Ahmed, T., Ali, N., & Rosenfeld, R. (2009). Speech vs. touch-tone: Telephony interfaces for information access by low literate users. In *Information and Communication Technologies and Development (ICTD), 2009 International Conference on* (pp. 447–457). IEEE.

Swaminathan, S., Medhi Thies, I., Mehta, D., Cutrell, E., Sharma, A., & Thies, W. (2019). Learn2Earn: Using mobile airtime incentives to bolster public awareness campaigns. *Proceedings of the ACM on Human-Computer Interaction, 3,* 1–20.

The World Factbook—Central Intelligence Agency. (2017). Retrieved from https://www.cia.gov/library/publications/the-world-factbook/fields/2103.html#136.

Thies, I. M., & others. (2015). User interface design for low-literate and novice users: Past, present and future. *Foundations and Trends® in Human–Computer Interaction*, *8*(1), 1–72.

Vashistha, A., Cutrell, E., Borriello, G., & Thies, W. (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* (pp. 417–426). ACM.

Vashistha, A., Sethi, P., & Anderson, R. (2017). Respeak: A voice-based, crowd-powered speech transcription system. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 1855–1866). ACM.

Vashistha, A., Sethi, P., & Anderson, R. (2018). BSpeak: An accessible crowdsourcing marketplace for low-income blind people. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM.

Vashistha, A., Garg, A., & Anderson, R. (2019). ReCall: Crowdsourcing on basic phones to financially sustain voice forums. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–13).

Vashistha, A., Saif, U., & Raza, A. A. (2019). The internet of the orals. *Communications of the ACM, 62*(11), 100–103.

Wang, H., & Singhal, A. (2018). Audience-centered discourses in communication and social change: The 'Voicebook'of Main Kuch Bhi Kar Sakti Hoon, an entertainment-education initiative in India. *Journal of Multicultural Discourses, 13*(2), 176–191.

White, J., Duggirala, M., Kummamuru, K., & Srivastava, S. (2012). Designing a voice-based employment exchange for rural India. In *Proceedings of the Fifth International Conference on Information and Communication Technologies and Development* (pp. 367–373). ACM.

Wolfe, N., Hong, J., Raza, A. A., Raj, B., & Rosenfeld, R. (2015). Rapid development of public health education systems in low-literacy multilingual environments: Combating ebola through voice messaging. In *ISCA Special Interest Group on Speech and Language Technology in Education (SLaTE)*. INTERSPEECH.

Zainudeen, A., Samarajiva, R., & Sivapragasam, N. (2010). Cellbazaar, a mobile-based e-marketplace: Success factors and potential for expansion.

Zue, V., Seneff, S., Glass, J. R., Polifroni, J., Pao, C., Hazen, T. J., & Hetherington, L. (2000). JUPlTER: A telephone-based conversational interface for weather information. *IEEE Transactions on Speech and Audio Processing, 8*(1), 85–96.