A Comparative Study of Speech and Dialed Input Voice Interfaces in Rural India

Neil Patel[1,2], Sheetal Agarwal[2], Nitendra Rajput[2], Amit Nanavati[2], Paresh Dave[3], Tapan S. Parikh[4]

[1]Stanford University HCI Group
Computer Science, Stanford, CA 94025
neilp@cs.stanford.edu

[2]IBM India Research Laboratory
New Delhi, India
{sheetaga, rnitendra, namit}@in.ibm.com

[3]Development Support Center
Ahmedabad, Gujarat, India
pdave68@gmail.com

[4]UC Berkeley School of Information
Berkeley, CA 94720
parikh@ischool.berkeley.edu

ABSTRACT

In this paper we present a study comparing speech and dialed input voice user interfaces for farmers in Gujarat, India. We ran a controlled, between-subjects experiment with 45 participants. We found that the task completion rates were significantly higher with dialed input, particularly for subjects under age 30 and those with less than an eighth grade education. Additionally, participants using dialed input demonstrated a significantly greater performance improvement from the first to final task, and reported less difficulty providing input to the system.

ACM Classification Keywords

H.5.2 User Interfaces: Voice I/O User Interfaces; H.5.2 User Interfaces: Evaluation; H.1.2 User/Machine Systems: Human Factors

Author Keywords

voice user interface, speech interface, isolated word, DTMF, India, rural development, semi-literate, ICTD

INTRODUCTION

Speech interfaces have been identified for their potential to increase access to information services in developing countries like India, where 480 million illiterate people reside [12]. Earlier research has demonstrated that automatic speech recognition (ASR) is possible for languages and dialects with limited speech resources, such as many of those spoken in India [10]. However, with this approach, acceptable error rates can only be obtained with a voice user interface (VUI) design that accepts a small number of distinct single word utterances at each node in the application (isolated word speech input).

Dual-tone multi-frequency (DTMF) is a mechanism for navigating voice user interfaces using the phone's numeric keypad. In this paper we present a study comparing isolated word speech and DTMF input VUIs for farmers in rural Gujarat, India. We conducted a controlled, between-subjects experiment with 45 participants, most of whom had less than an eighth grade education. The goal of our study was to compare performance and user preference between the two input modalities and to correlate the results to users' education levels and age. Our results show that DTMF outperformed speech in terms of task completion rate and learnability, and users reported significantly less difficulty providing input using DTMF.

RELATED WORK

Many studies comparing input modalities for VUIs have been conducted in developed countries [3, 4]. Lee and Lai compared a dial interface to a fully functioning natural language system. They found that user preference depends on the task being completed — DTMF was preferred for linear tasks (i.e. listening to voicemails in the order received), while speech was preferred for non-linear tasks (i.e. listening to voicemails from a specific acquaintance in random order) [6]. Delogu et. al. compared DTMF to three different speech input systems and found no difference in performance, but found a user preference for DTMF over an isolated word interface [2]. In this paper we report results from an important user population outside the scope of these studies. Our experiment involved farmers from rural Gujarat, a state located in western India, where the native language is Gujarati. 87% of the participants had never used a computer and 73% of the participants had less than an eighth grade education.

Other researchers have investigated the design of VUIs for such populations. Sherwani developed a VUI in Urdu for semi-literate community health workers in Pakistan [11]. Plauche designed a VUI in Tamil for access-

CHI 2009, April 4 - 9, 2009, Boston, Massachusetts, USA.

ing agricultural market information [10]. She demonstrated that by restricting the input vocabulary to 2-3 words per node, a VUI using only 15 speakers' speech data could achieve an error rate of 2% or less. The tradeoff for accuracy was that most prompts were yes-or-no questions. In this study, we evaluated a system which uses a viable alternative strategy for limited resource languages: using a recognizer trained on another language with copious speech resources (English in this case).

Prior research has pointed out that numerical literacy can be leveraged for designing user interfaces accessible to semi-literate users [8]. Several researchers have experimented with DTMF interfaces in developing regions and found them preferable to speech input for women users, and in situations where speaking out loud could raise privacy concerns [7, 9]. However, we are not aware of other published studies directly comparing these two input modalities for the population we are considering - users with limited education and experience with computer interfaces.

PROTOTYPE

For our study, we designed *Avaaj Otalo* ("voice-based community forum"), a Gujarati language application allowing farmers to access agricultural information over the phone. To accommodate novice users, our main design goal for the interface was simplicity. Functionality was laid out in hierarchical menus, and all tasks were linear. We limited all navigational nodes in the application to two or three options. To avoid command ambiguity, only directive-style prompts were used, telling the user specifically what commands they could give.

We partnered with Development Support Center (DSC), an NGO in Ahmedabad, Gujarat, to conduct a joint needs-finding exercise, based on which three system features were identified and implemented. The *announcement board* is a list of headline-like informational snippets, uploaded to Avaaj Otalo by DSC staff or other agriculture experts several times per week. The *radio archive* lets the caller listen to archived radio programs produced by DSC on agricultural topics of current interest. Finally, Avaaj Otalo allows farmers to record their own *questions*, for review and response by experts.

We implemented both isolated word speech and DTMF versions of Avaaj Otalo. Prompts were recorded in a professional studio by one of the DSC radio program's popular female voice personalities. Barge-in input was disallowed for both treatments. Figure 1 shows a sample dialog with Avaaj Otalo.

Avaaj Otalo was built and deployed using IBM Research India's WWTW [5] platform. For the speech recognition, Gujarati commands were converted to lexicons using the American English phoneme set. In our experiment, the system performed with a recognition accuracy of 94%. Although this is lower than Plauche's Tamil system (98% accuracy), the difference reflects the cost of a larger command vocabulary for limited resource languages.

AO: Welcome to Avaaj Otalo! You can get to information by saying a single word. To ask a question, say 'question'; to listen to announcements, say 'announcements'; to listen to the radio program, say 'radio'.
User*: I want to ask a question.*
AO: Sorry, I didn't understand. I can only understand single words. Do you want to ask a question... yes or no?
User*: Yes*
AO: OK, you want to ask a question. To ask a question about agriculture, say 'agriculture'; for animal husbandry, say 'animal'. ....
AO: OK, you want to ask a question about pests in cotton. Please say your question slowly and clearly after the beep.
User*: How can I protect my cotton crop from mili bugs?*

Figure 1. A sample interaction with Avaaj Otalo. The DTMF version of the application had identical prompts except that command options were mapped to numeric keys.

EXPERIMENT

We tested Avaaj Otalo with 45 participants recruited from ten districts throughout rural Gujarat. To participate, we only required that subjects be farmers by profession. We focused on recruiting small-scale farmers; the median farm size was 10 acres. All of the participants spoke Gujarati as their primary language, and none spoke English. The majority of participants (87%) reported never having used a PC.

We did not use a within-subjects experiment design because we felt the simplicity of the application would have introduced a priming effect. Input modality (speech vs. DTMF) was randomly assigned to each user, but was anonymously corrected to maintain balance across age, education and gender.

Testing sessions were led by a DSC staff member who had experience communicating with the target user group. Participants were first introduced to the system and its features, and were assured that it was the system that was being tested, not them. Each participant completed three tasks with Avaaj Otalo corresponding to its three features (listening to announcements, listening to archived radio program recordings, and posting questions), ordered by increasing difficulty.

We designed Avaaj Otalo to be responsive to input errors. If the system could not recognize user input, or if the user was silent, a follow-up prompt would ask the user to try again. If input was again not recognized, the system reverted to a series of yes-or-no prompts, offering each option serially. We classified a task as failed if the user either navigated to a part of the application that was not called for by the task, or failed to get passed the yes-or-no prompts after several attempts with no sign of recovery.

We tested 38 participants in a quiet office, with only the DSC staffer and two researchers as observers. We used

a landline phone in both treatments. The remaining 7 participants, all women, were tested in their homes due to their traveling constraints. In the field, we attempted to be faithful to the office environment by testing in a quiet room with only the researchers and one family member of the participant present. A landline phone was not available, so we used a mobile phone. Participants in the DTMF treatment were provided a headset so that the dialpad could remain in front of them (see figure 2).

Figure 2. Testing the DTMF interface with a participant at her home.

Capturing Data

We used several methods to record experimental data. For collecting demographic information, we administered a pre-test questionnaire. For performance measures, we instrumented our prototype to log task completion, errors, and call duration. During the test, two researchers noted points of difficulty, facial expressions, and comments made during the call. To measure user satisfaction, ease of use, and learnability, we administered a post-test questionnaire with Likert scales.

RESULTS

Performance Results

The overall task completion rate with DTMF was significantly higher than with speech (74% vs. 61%; $p < 0.05$). Figure 3 shows the breakdown by task, and according to age and education level. The third task, recording a question, consisted of three subtasks: categorizing the question, recording the question, and recording the participant's name and location. Categorization (task 3a) was the most difficult because it required traversing several levels, choosing one of nine crops, and one of six agricultural topics. For this subtask, DTMF users had a significantly better completion rate than speech (the completion rates were also better for the other two subtasks, but not significantly so).

Participants using the DTMF interface also demonstrated a significantly greater performance improvement between the first and third task. We calculated the effect size using Cohen's d repeated measures analysis, corrected for correlated datasets [1]. DTMF users experienced a

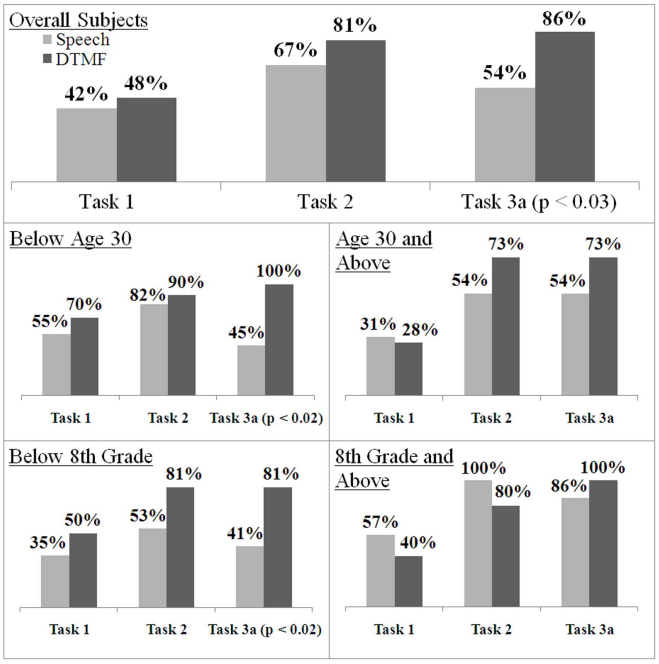

Figure 3. Task completion rates for speech (light gray) and DTMF (dark gray) versions. P-values are given where rate differences were significant.

	Task1	Task2	Task3
DTMF	48%	19%	29%
Speech	63%	42%	42%

Table 1. Percentage of users who reported each task as either "difficult" or "very difficult".

"large positive difference" (Cohen's d-value = 0.99) in completion rates between task 1 and 3. With speech the effect was a "small positive difference" (Cohen's d-value = 0.26).

Despite the difference in task completion rate, there was no significant difference in user satisfaction. In both groups, over 80% of users reported that they found it easy to access information from the system. Over 75% of both groups said they would "definitely" use such an application if it was made available.

User Perception of Difficulty

Table 1 displays the percentage of users who reported that a particular task was either "difficult" or "very difficult", based on a five-point Likert scale. Across all tasks, the percentage of such responses was 49% for speech and 30% for DTMF ($p < 0.05$). When specifically asked whether they faced any difficulty providing input to the system, 81% of DTMF users answered "no" or "definitely no", compared to 38% for speech users ($p < 0.01$).

DISCUSSION

Our most consistent result was the success of dialed input relative to speech, confirming results obtained

in other settings [2, 6]. Our observations indicated two main reasons why speech input was less successful. First, users expressed discomfort speaking single word commands, which was perceived as unnatural. "Talking to the computer" was an unfamiliar idea; DTMF users may have had an easier time forming a mental model of the system. The second reason was difficulty in recovering from errors made by either the system (recognition error) or the user (bad or no input). With speech input, the task completion rate was 42% when one or more recognition errors occurred, compared to 67% when no errors occurred ($p < 0.05$). Given the recent emphasis on designing limited vocabulary speech interfaces for semi-literate users, it is notable that the only group who performed better using speech for multiple tasks was the most educated group. This indicates that less educated users may have more difficulty recovering from recognition errors.

Due to the difficulty and expense of providing training, an interface that is easy to learn and understand is a key design consideration for information services serving remote populations. No users expressed difficulty in understanding how to operate the system through dialed input, including several fully illiterate participants. However, one difficulty with the DTMF interface was in transitioning between dialed input and speaking, which was required in the final task for recording the user's question and personal information. A difficulty across both modalities was navigating command-driven menus and knowing when to provide input. Every spoken prompt was followed by a beep to indicate that input was requested. The prompts did not explicitly mention the beep, and many users either gave input too early or not at all.

Difficulties notwithstanding, the participants' response to the application was unanimously enthusiastic. Many farmers told us that the ability to access information at any time would have a significant impact on their farming practices. A few farmers singled out the ability to share their personal experiences with other farmers and with DSC staff as a key benefit of the system.

The main limitation of the study is its external validity. The study was conducted in optimal conditions for both accurate speech recognition (a calm, quiet environment) and easy dialing (placing the dialpad in front of users). A real-world deployment must support usage in a diverse range of scenarios. We plan to conduct a more realistic assessment of the usage and impact of this system after it is deployed across Gujarat. The study's generalizability is also limited by the narrowness of the type of task that was tested. Linear tasks with low perplexity are amenable to DTMF input, and it is possible that speech input could outperform DTMF in more complex scenarios.

CONCLUSION

In this paper, we presented a comparative study of speech and dialed input for a user population with limited literacy, familiarity with technology, and for a language with limited speech resources. We developed *Avaaj Otalo*, an application for farmers to access relevant and timely agricultural information. We found that dialed input outperforms speech, both in terms of task completion rate and users' perception of difficulty. We plan on deploying Avaaj Otalo for access throughout Gujarat next year.

ACKNOWLEDGMENTS

This work was supported by IBM Research India and the Stanford School of Engineering. The authors thank Development Support Center, Arun Kumar, Anupam Jain, and Priyanka Manwani for their contributions. A special thanks to Scott Klemmer for his invaluable guidance. We are sincerely grateful to the farmers who helped us design and test Avaaj Otalo.

REFERENCES

1. J. Cohen. *Statistical power analysis for behavioral sciences*. Lawrence Erlbaum Associates, 1998.
2. C. Delogu, A. D. Carlo, P. Rotundi, and D. Sartori. A comparison between DTMF and ASR IVR services through objective and sujective evaluation. In *IVTTA*, 1998.
3. J. Foster, F. McInnes, M. Jack, S. Love, R. Dutton, and I. Nairn. An experimental evaluation of preference for data entry method in automated telephone services. In *Behavior and Information Technology*, 1998.
4. M. Goldstein, I. Bretan, E. L. Sallnas, and H. Bjork. Navigational abilities in voice-controlled dialogue structures. In *Behavior and Information Technology*, 1999.
5. A. Kumar, N. Rajput, D. Chakraborty, S. Agarwal, and A. Nanavati. WWTW: The World Wide Telecom Web. In *SIGCOMM Workshop on Networked Systems for Developing Regions*, Japan, Nov 2007.
6. K. M. Lee and J. Lai. Speech versus touch: A comparitive study of the use of speech and dtmf keypad for navigation. *International Journal of Human-Computer Interaction*, 2005.
7. T. J. Ndwe. personal communication, 2008.
8. T. S. Parikh, K. Ghosh, A. Chavan, P. Syal, and S. Arora. Design studies for a financial management system for micro-credit groups in rural india. In *ACM Conference on Universal Usability*, 2003.
9. M. Plauche. personal communication, 2008.
10. M. Plauche, U. Nallasamy, J. Pal, C. Wooters, and D. Ramachandran. Speech recognition for illiterate access to information and technology. In *International Conference on Information and Communications Technologies and Development*, 2006.
11. J. Sherwani, N. Ali, S. Mirza, A. Fatma, Y. Memon, M. Karim, R. Tongia, and R. Rosenfeld. Healthline: Speech-based access to health information by low-literate users. In *International Conference on Information and Communications Technologies and Development*, 2007.
12. United Nations Development Program. Human development report, 2004.