VoicePedia: Towards Speech-based Access to Unstructured Information

J Sherwani [1], *Dong Yu* [2], *Tim Paek* [2], *Mary Czerwinski* [2], *Y C Ju* [2], *Alex Acero* [2]

[1] Carnegie Mellon University, Pittsburgh, PA, USA
[2] Microsoft Research, Redmond, WA, USA

jsherwan@cs.cmu.edu, {dongyu,timpaek,marycz,yuncj,alexac}@microsoft.com

Abstract

Currently there are no dialog systems that enable purely voice-based access to the unstructured information on websites such as Wikipedia. Such systems could be revolutionary for non-literate users in the developing world. To investigate interface issues in such a system, we developed VoicePedia, a telephone-based dialog system for searching and browsing Wikipedia. In this paper, we present the system, as well as a user study comparing the use of VoicePedia to SmartPedia, a Smartphone GUI-based alternative. Keyword entry through the voice interface was significantly faster, while search result navigation, and page browsing were significantly slower. Although users preferred the GUI-based interface, task success rates between both systems were comparable – a promising result for regions where Smartphones and data plans are not viable.

Index Terms: dialog system, information access

1. Introduction

For users in emerging regions, where poverty and illiteracy often make desktop computing unaffordable and unusable, telephone-based spoken dialog systems offer a potentially viable alternative. Dialog systems for information access by non-literate users are being explored in agriculture [1], health [2], and even dialog system generation [3]. While dialog systems for information access have been heavily researched in both academia [4] and industry (e.g., Amtrak's Julie, TellMe's 1-800-555-TELL, etc.), the focus has been on highly structured information such as weather, directory assistance, movie show-times, and stock quotes. There has been little work done on dialog systems for accessing semi-structured or unstructured information, such as that found on the web, as literate consumers (who have been the end-users of most computer interface design) would more effectively access these through graphical, non-speech modalities.

Many spoken dialog systems (such as [5]) scrape the web and parse out information to fill a structured database, while question-answering dialog systems such as [6] rely on a corpus of knowledge that is highly semantically structured. However, these systems either only cover a very specific domain (e.g. restaurants), or depend on semantic parsing techniques that do not cover all possible questions and answers, or involve heavy preprocessing of the information that is to be extracted. Specifically, there are no telephone/voice systems that enable purely spoken access to informational search [7]. Such systems could potentially enable non-literate users to access digital libraries which have hitherto been inaccessible [8]. To address this gap, we introduce VoicePedia, a telephone-based dialog system for searching and browsing the Wikipedia website.

This paper is divided into two sections. First, we elaborate on the VoicePedia system and its parameter space. Second, we describe the preliminary results of a usability study in which participants engaged in a search task with two distinct parameter settings of VoicePedia, as well as with a standard Smartphone browser. We report both quantitative and qualitative measures, and discuss implications for the usefulness of similar voice interfaces.

2. VoicePedia

The VoicePedia interface was designed to mimic the web search experience, site-constrained to the Wikipedia domain. The interaction consists of three major dialog phases: a) keyword entry, b) search result navigation, and c) web-page navigation.

In the keyword entry phase, the user is asked to say search keywords one at a time. After each keyword, the system uses either of two disambiguation strategies – Best Guess or Web Suggest – to choose a hypothesis from the n-best list, and echoes this hypothesis to the user, who may then either correct it, or continue with the next keyword. Once all keywords have been entered, the user says "that's all" to move to the next phase.

Based on the search keywords, the system performs a web search query of the form: **site:en.wikipedia.org k_1 k_2 k_3… k_n** where k_i represents the *i*th keyword. The user is then presented with a list of the page titles of the top 10 search results, with pauses after the 3rd and 6th results. The user is able to select a search result by repeating the title, or by saying the search result number, or by saying "that one" after hearing a specific search result, or by asking for the first or last search result.

After selecting a search result, the system fetches the Wikipedia web-page corresponding to the search result, and parses it to extract page sections. Most Wikipedia pages have an untitled introductory section, followed by titled sections which are listed in the table of contents. VoicePedia creates its own table of contents for each page by starting off with a section titled "Introduction" which consists of the otherwise-untitled introductory section, and adds the other titled sections from the table of contents. The user is given a prompt to choose which section they would like to hear read aloud, followed by the list of all section titles, with a pause after each title. After choosing a section title, VoicePedia tells the user how many paragraphs there are in that section, and begins reading from the first paragraph. After each paragraph is crossed, a short earcon is played to denote the end of the paragraph. The user is able navigate between paragraphs and sentences via appropriate voice commands.

Additionally, while navigating the Wikipedia page through VoicePedia, the user can ask to search for a specific word in the page. VoicePedia reads aloud the line at which the word occurs, and the user can navigate between the next and previous search result with appropriate voice commands.

2.1. Implementation

The VoicePedia system has been implemented in C# in the Microsoft Speech Server environment. Unigram grammars of similar size have been found to be the optimal choice for web search keyword entry [9], and so for the keyword entry phase, we used a generic 100K word unigram grammar (generated from WSJ text). While it would have been optimal to bias the unigram towards Wikipedia searches, we wanted to test the system in the absence of query log data that is often proprietary. For our experiment, we made sure that the unigram's coverage was sufficient for the given stimuli. The Microsoft Live Search engine was used to perform web queries.

2.2. Disambiguation Strategies

In the keyword entry phase, the system is given a parameter to use either the Best Guess or Web Suggest disambiguation strategies to choose between entries in the n-best list. In the Best Guess strategy, the system chooses the hypothesis with the highest confidence. If this confidence is less than 0.5, it is treated as a non-understanding.

Early on in the development of VoicePedia, it was evident that the lack of semantic information in the baseline Best Guess method was hurting system performance. For instance, if the user said "Tom Cruise", the system was very likely to go with "com cruise" or "palm cruise" as a direct result of the unigram language model. The Web Suggest strategy was devised to add semantic information into the decision of which n-best hypothesis to choose. In this strategy, the system goes through the n-best list, and for each entry, regardless of the confidence level, sends it to a search keyword suggestion engine (such as Google Suggest) and retrieves the top ranked suggested completion, as well as the number of search results for the top ranked suggestion. If an entry from the n-best list gets zero suggestions, it is removed from consideration. The processed n-best list is then sorted by number of search results, and presented to the user to disambiguate. For example, if the user says **sierra** with an n-best list of **sierra**, **cia**, and **ciera**, the system would ask the user "did you say sierra as in sierra trading post, cia as in cia factbook, or ciera as in ciera lyrics?" After this disambiguation dialog, the system continues with implicitly confirming the user's choice.

2.3. SmartPedia

Since we were interested in investigating the difference between using a voice-only modality versus a handheld mobile device with a limited GUI (which is the primary modality for informational searches in mobile situations), we implemented a SmartPhone version of VoicePedia, called SmartPedia. This system uses the same search query format, the same search engine, and the same Wikipedia parser as in the VoicePedia system, but was designed for a web-browser on a SmartPhone. Since VoicePedia does not indicate any hyperlinks or images within a Wikipedia entry, the SmartPedia interface strips these out. For keyword entry, both T9 and multi-tap text entry methods were made available.

3. Related Research

[7] defines three classifications of web search queries: informational, navigational and transactional. In informational search queries, the intent of the user is to access information assumed to be present on one or more web pages, e.g., queries such as "cars", "San Francisco", "normocytic anemia" and "Scoville heat units". While many web-sites effectively deal with such queries, there are no known spoken dialog systems that enable purely voice-based access to such search engines.

Various dialog systems [4] and [5] enable voice-based access to domain-specific information scraped from the web. However, each of these systems focuses on specific domains (such as restaurants, flight schedules, weather).

Question-answering systems, both spoken and text-based [6] focus on presenting the user with one or more potential answers to the user's question – removing the need for users to transform their questions into a query, and to then navigate search results and webpages to find the answer. However, these focus on providing the user with a short nugget of information, and are not able to give detailed instructions that the user can navigate as in a web-page.

There has been work on voice-based web search [9]. However, the voice interface focused exclusively on the keyword entry phase, and required access to a desktop web browser to view results and browse web-pages, and so was not a purely voice-based interface.

4. Experiment

To assess the usability of the VoicePedia system, we conducted a controlled experiment in which participants engaged in an informational search task using a list of trivia questions as stimuli. For both VoicePedia and SmartPedia, participants were given a short tutorial of each system before they used it to find answers to the trivia questions. The order of presentation and the choice of disambiguation strategy were counterbalanced among participants.

4.1. Method

4.1.1. Procedure

Participants were presented with a web interface which was used throughout the study. Each webpage contained some text, along with input fields to get user feedback. Every page contained buttons marked 'Next' and 'Previous' which participants were instructed to use to go through the study. There were a total of 108 pages that each participant went through, covering tutorials, trivia question-answering, and subjective feedback elicitation.

Participants started off with a brief overview of the entire experiment, and then were given a short tutorial for the first system (either VoicePedia or SmartPedia). This was followed by seven trivia questions that were to be answered using that system. Participants were then given a number of questions to rate their experience with the system on a 7-point Likert scale. After this, the process was repeated for the second system (tutorial, trivia questions, feedback). Finally, participants were given a set of questions comparing the two systems on various factors.

4.1.2. Stimuli

The 14 trivia questions, along with the associated search keywords, were chosen to ensure that all keywords were within the unigram grammar of the system, and that the answers could be found using the top 10 search results from Wikipedia. Five of the questions had the correct answer as one of the search engine results (and so could be answered after finishing 2 out of the 3 interface phases). Each question was displayed on a webpage, along with 4 possible answers from which participants had to select one.

4.1.3. Design

The experiment followed a 2 (System) x 2 (Disambiguation) factorial design. System was within subjects, and disambiguation was between subjects. Question answering task completion time was the main dependent variable. In addition, we also looked at answer correctness, time spent in each phase, and qualitative questionnaire ratings. Presentation of trivia questions was blocked by interface. Participants completed 1 practice task and 7 experimental tasks in each System condition. Order of presentation of the system conditions was counterbalanced across participants.

4.1.4. Participants

The participants were 10 employees of Microsoft. All were native speakers of American English, and had extensive experience of using web search interfaces.

5. Results

5.1. Objective Metrics

Task success was defined as the number of correctly answered trivia questions. Although task success was slightly higher for VoicePedia, when we performed a one-way ANOVA on task success, we found no main effect for System. This suggests that the two Systems may be equivalent in terms of their effectiveness. This result was surprising given that SmartPedia utilizes a visual user interface and VoicePedia is voice only. We discuss this further later.

Wall-clock task completion time was defined as the total amount of time from receiving a trivia question to clicking the "Next" button after choosing one of the 5 options in the radio button list. However, because of a slight timing error, wall-clock time for any given task was much higher than the actual time taken to perform the task. Thus, we defined adjusted task completion time as the actual time taken to complete the task. Also, in 5 out of 70 of the VoicePedia tasks, the speech recognition engine was unable to recognize the same keyword after 5 repeated utterances (due to non-word entries in the unigram), and these dialogs were removed from the analysis.

With the inflated wall-clock task completion time, we performed a one-way ANOVA, and found a main effect for System ($F(1,132)=12.70$, $p<.01$). However, a one-way ANOVA on adjusted task completion time revealed no main effects.

Additionally, we calculated the time taken on each of the three phases (keyword entry, search results navigation, and page navigation).

Speech has been known to be faster than text entry on mobile devices [10]. Our study corroborated these results – while time spent on the search and page navigation phases was higher in the VoicePedia system ($F(1,132)=4.95$, $p<0.05$, and $F(1,132)=17.67$, $p<0.001$, respectively), it was lower in the keyword entry phase. Although this may have been due to the fact that in the SmartPedia system, the search query was erased whenever the page was reloaded, and so any attempts to edit the query to refine search results meant that all keywords had to be re-entered whereas this was not the case in the VoicePedia system, such keyword re-entry only occurred in 5 out of the 70 SmartPedia tasks; hence, it is unlikely to have accounted for the large significant difference ($F(1,132)=6.09$, $p<0.05$).

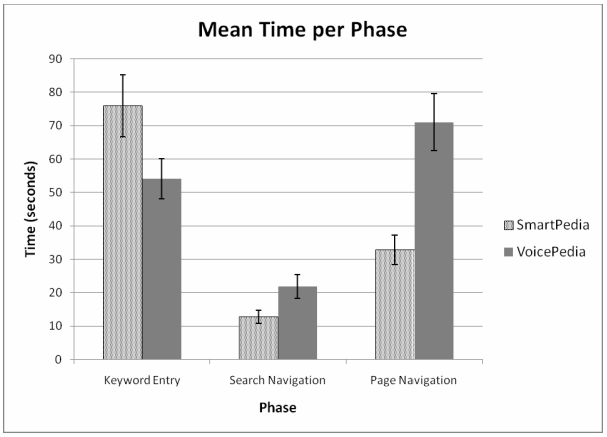

Fig. 1: *Mean time per phase. Error bars indicate standard error. All differences for each phases were statistically significant ($p < .05$).*

Finally, there was no significant difference in phase completion time across disambiguation strategies. Of the 364 turns for keyword entry, the correct answer was the best hypothesis in 58% of all turns, while it was in the n-best list in 86%. 59% of all turns involved disambiguation, of which 55% were Best Guess, while 45% were in the Web Suggest case. The Web Suggest strategy chose a unique hypothesis without needing user disambiguation in 21% of the cases. In 62% of the remaining cases, however, the Best Guess strategy would have chosen the correct hypothesis automatically. Thus, it appears that the high cost of the Web Suggest (verbose prompts, higher cognitive load) was greater than the benefits (considering hypotheses that might have been less confident but correct).

5.2. Subjective Metrics

Three sets of subjective metrics were measured in the study. In all three sets, participants were asked to rate the systems using 7-point Likert scales on a series of questions, many of which were based on the SASSI framework [11].

In the first two sets, participants were given the same series of questions pertaining to a system immediately after using it. A one-way ANOVA on their responses revealed a main effect for System on learnability ($F(1,18)=12.13$, $p<.01$), cognitive load ($F(1,18)=4.6$, $p<.05$), satisfaction with the system for the given tasks ($F(1,18)=7.13$, $p<.05$), understandability of search results ($F(1,18)=16.35$, $p<.001$), and difficulty in remembering interface options while browsing search results ($F(1,18)=24.05$, $p<.001$). These results are shown in Figure 2.

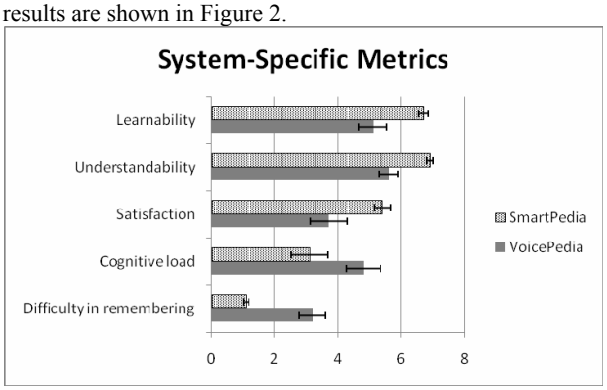

Fig 2: *User ratings for VoicePedia and SmartPedia. Error bars indicate standard error. All differences were statistically significant ($p < .05$)*

In the second set of ratings, participants were asked to rate both systems relative to each other at the end of the study, once they had completed all tasks on both systems. In general, participants favored SmartPedia, except when asked which system was faster in entering keywords.

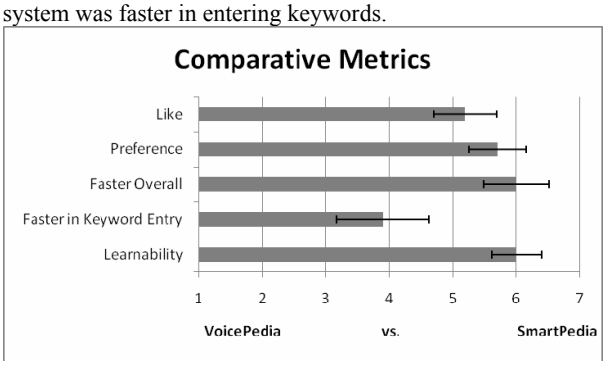

Fig 3: *User ratings of VoicePedia and SmartPedia. Error bars indicate standard error.*

Additionally, participants were asked to describe their experience and preferences textually. 9 out of 10 participants reported issues with the quality of speech recognition, while 5 expressed dissatisfaction with the text-to-speech voice's intelligibility. Two said that the cognitive load associated with the task itself was too high for the VoicePedia system.

Half of the participants said that they would have liked a multimodal system that enabled speech on the Smartphone interface.

6. Discussion

The above results show that task success is comparable in both systems, yet participants preferred the GUI-based system. This can be explained by the qualitative metrics, which show strong differences in both cognitive load, and in the perceived complexity of the interface. It is interesting to note that even though the voice interface was designed to mimic the GUI interface as closely as possible, speech recognition and synthesis errors, along with the extra turns for confirmations, and the difficulty in "browsing" lists and large amounts of spoken text made SmartPedia the preferred system.

One potential way to improve VoicePedia performance would be to modify the search engine to detect and offer corrections for potentially misrecognized words (analogous to how current search engine behavior with misspelled words).

Even if speech recognition and synthesis errors are reduced, however, issues related to cognitive load and perceived complexity are likely to remain, as a direct result of the opacity of speech interfaces. Solutions to these issues necessarily require interface-level improvements, which should start by reworking the interface from the ground up for a voice-only modality (rather than mimicking an existing GUI interface). One promising direction is Schneiderman's mantra of a unified interface that provides "overview, zoom, filter, details" [12] – the web search process can be made to quite neatly fit into this framework, and the unification of the various phases of web search may make it simpler for users. Other possibilities for improvement include dynamically altering the verbosity of prompts based on expected utility and/or cost for the user.

7. Conclusions

In this paper, we introduced VoicePedia, a purely voice-based interface for informational search over the Wikipedia website. We presented results from a user study comparing the use of VoicePedia with SmartPedia – a Smartphone GUI-based interface for the same purpose. We showed that keyword entry was faster through speech, although search navigation and page navigation were slower. Even though task success was comparable across both systems, users overwhelmingly preferred the GUI interface, based on speech recognition and synthesis issues, as well as cognitive load and complexity issues with the speech interface.

Testing such systems with our intended audience – non-literate users in the developing world – is an active area of research [2]. Since web search is not a familiar interface metaphor in such users, there may be no advantage of mimicking a GUI interface, and it would be prudent to understand what interface metaphors may work in their context. On the other hand, since non-literate users cannot type keywords or browse search results and web pages, VoicePedia would clearly be preferable to SmartPedia for such users. In either case, our user study indicates that it is possible for an interface such as VoicePedia to be effectively used for information access, and empirical evidence is now needed to see whether or not this would hold true for non-literate users in emerging regions.

8. References

[1] Plauche, M., Nallasamy, U., Pal, U., Wooters, C., and Ramachandran, D., "Speech Recognition for Illiterate Access to Information and Technology", Proc. Int'l Conf. on ICTs and Development, 2006.

[2] Sherwani, J., "Are Spoken Dialog Systems Viable for Under-served Semi-literate Populations?", Ph.D. proposal, http://www.cs.cmu.edu/~jsherwan/proposal.pdf

[3] Kumar, A., Rajput, N., Chakraborty, D., Jindal, S., and Nanavati, A., "VOIGEN: A Technology for Enabling Data Services in Developing Regions", IBM Tech Paper.

[4] Rudnicky, A., Thayer, E., Constantinides, P., Tchou, C., Stern, R., Lenzo, K., Xu, W., Oh, A., "Creating natural dialogs in the Carnegie Mellon Communicator System", in Proceedings of Eurospeech, 1999.

[5] Gruenstein, A., Seneff, S., and Wang, C., "Scalable and Portable Web-Based Multimodal Dialogue Interaction with Geographical Databases". Proc. Interspeech 2006.

[6] Schofield, E., and Zheng, Z., "A speech interface for open-domain question-answering", in Proc. ACL 2003.

[7] Broder, A., "A Taxonomy of Web Search", in Proc. ACM SIGIR 2002.

[8] Deo, S., Nichols, D., Cunningham, S., Witten, I., "Digital Library Access for Illiterate Users". Proc. Int'l Research Conf. on Innovations in IT 2004.

[9] Franz, A., Milch, B., "Searching the Web by Voice", http://people.csail.mit.edu/milch/papers/gvs.pdf

[10] Lewis, J.R., and Commarford, P.M. "Developing a Voice-spelling alphabet for PDAs", IBM Systems Journal, Vol. 42, No. 4, pp. 624-638, 2003.

[11] Hone, K. and Graham, R. "Subjective Assessment of Speech-System Interface Usability". Proc. Eurospeech 2001.

[12] Shneiderman, B., "The Eyes Have It: A Task by Data Type Taxonomy For Information Visualizations", Proc. IEEE Symposium on Visual Languages 1996.