The Benefits of Augmenting Telephone Voice Menu Navigation with Visual Browsing and Search

Min Yin
IBM Almaden Research Center
650 Harry Road, San Jose, California, USA
myin@us.ibm.com

Shumin Zhai
IBM Almaden Research Center
650 Harry Road, San Jose, California, USA
zhai@almaden.ibm.com

ABSTRACT
Automatic interactive voice response (IVR) based telephone routing has long been recognized as a frustrating interaction experience. This paper presents a series of experiments examining the benefits of augmenting telephone voice menus with coordinated visual displays and keyword search. The first experiment qualitatively studied callers' experience of having a visual menu on a screen in synchronization with the telephone voice menu tree navigation. The second experiment quantitatively measured callers' performance in time and accuracy with and without visual display augmentation. The third experiment tested keyword search in comparison to visual browsing of telephone menu trees. Study participants uniformly and enthusiastically liked the visual augmentation of voice menus. On average with visual augmentation callers could navigate phone trees 36% faster with 75% fewer errors, and made choices ahead of the voice menu over 60% of the time. Search vs. browsing had similar navigation performance but offered different and complementary user experiences. Overall our studies conclude that telephone voice menu navigation can be significantly improved with a visual channel augmentation, resulting in both business cost reduction and user experience satisfaction.

Author Keywords
voice menu, instant messaging, telephone, multi-modal interaction, integrated user experience, keyword search, visual manual browsing.

ACM Classification Keywords
H.5.1 Multimedia Information Systems. H5.2 [Information interfaces and presentation]: User Interfaces. - Graphical user interfaces, Interaction styles.

INTRODUCTION
Often referred as "touchtone hell", the difficulty and frustration with automatic interactive voice response (IVR) based phone call routing can be experienced first hand when one tries to reach the right human agent through the telephone lines of corporations, financial institutions, technical support centers, hospitals, airlines, and government agencies. The problems of dealing with IVR systems are also documented in the HCI research literature which has long recognized that "The current generation of telephone interfaces is frustrating to use, in part because callers have to wait through the recitation of long prompts in order to find the options that interest them." [8]. Researchers have studied how to better design the voice menu to ease the caller's frustration [e.g.5, 7, 10]. For example, Suhm, Freeman and Getty found that long touch tone menus route the caller more efficiently than short menus, since long menus reduce the number of menu layers to navigate [10]. Others, however, suggest one way of easing the limitations of auditory menus is to employ greater depth in the hierarchy and "reap the benefits of funneling and insulation" [7]. Inspired by people's ability to shift their gaze in order to skip uninteresting items and scan through large pieces of text, Resnick and Virzi proposed "skip and scan" as an alternative touchtone interface style in which callers issue explicit commands to accomplish skipping actions [8].

Despite these efforts, the same voice menu based IVR remains the state of the art. The difficulty of navigating voice menus is fundamentally rooted in the nature of auditory information. Sound expands in space but localizes in time. Consequently, unlike graphical and textual menus, voice menus are sequential at a fixed pace, either too fast (when the information is critical) or too slow (when the information is uninteresting) for the caller. A long voice menu is frustrating to the caller since it requires the caller to memorize many choices in order to compare and select the most reasonable one. Short and broad categories can also be difficult because the caller is often unsure which category will lead to the desired end. It is often difficult to tell if a particular category of functions suits the caller's need until choices at a lower level of the hierarchical menu are heard.

If the caller is impatient and fails to catch, or forgets, a particular choice, he or she often has to start all over. In contrast, visually scanning and choosing from a menu displayed with text can be done at the user's own pace. One can scan and compare the menu items back and forth without having to commit them to memory. One can also more easily jump between different levels of a visually presented hierarchical menu structure.

The idea of visually displaying the voice menu in IVR systems on a screen to the caller has been proposed many times in the invention disclosure literature. For example, Kreitzer [3] describes the concept of displaying the text of the voice menus onto a screen built into the phone set, together with handshaking mechanisms between the IVR and the caller's telephone. Similar proposals have also been disclosed by Fawcett, Blomfield-Brown and Strom [1], Hillier [2], and Narayanaswami [6]. On a related topic, Whittaker and colleagues at the AT&T labs have also explored creating a visual analogue of speech data from voice mails to support visual scanning, search and information extraction [11]. Recently, Yin and Zhai [12] proposed FonePal, a solution helps callers navigate telephone voice menu by automatically launching a coordinated visual channel for the caller. FonePal uses a "cross-device user experience integration" approach to provide the visual support. After one time ID (e.g. caller ID and instant messaging ID) registration, when a person makes a call to an IVR system from a phone, a FonePal system automatically delivers a graphical menu corresponding to the IVR voice menu through the Internet (instant messaging) onto a computer that is, as suggested by the IM client status, being actively used by the caller. As the caller selects the desired choices either by pressing the phone keypad or by clicking on the graphical menus on the computer screen, both voice and visual information are updated accordingly. Yin and Zhai described the motivation, design rationale, system architecture and various implementations of the FonePal solution in [12] .

Surprisingly, although the relative difficulty of using voice menus is generally recognized in the HCI literature such as the Handbook of Human-Computer Interaction [5, 9], no comparative study on the effect of visually augmenting voice menus could be found. Without careful empirical investigation, there are many conceptual and practical questions to be answered: How much benefit is there to visual augmentation of voice menus? Is the benefit worthwhile? Would the simultaneous display of auditory and visual information be mutually distracting? Which channel do users rely on more if both are present? How much would the users like it and why? Would keyword search be helpful to phone tree navigation? We therefore carried out a series of three empirical studies focusing on: 1. The subjective caller experience of navigating visually augmented voice menu. 2. Quantitative performance improvements that can be brought by visual augmentation to voice menus. 3. Search vs. browsing as a means to visually augmented phone tree navigation. To keep the experimental scope manageable while still addressing the most important concerns in the real world we decided to focus on comparing with and without visual augmentation and keep the voice menu intact. A pure visual condition may also be considered in a future study.

EXPERIMENT 1 – QUALITATIVE AND SUBJECTIVE EXPERIENCE

The first experiment focused on the qualitative and subjective aspects of using visually augmented voice menus.

Design and methodology

Experimental system: For this study, we used FonePal as proposed in [12] to visually display menu tree to the participant. We also implemented a complete IVR system whose content and structure were copies of the technical helpline of the IBM Corporation (1-888-IBM-HELP). We choose the content and structure of a real IVR system to induce realistic user reactions and experiences at an appropriate level of task complexity. Corporate helpline is a type of service people often deal with in the workplace and is also where FonePal is most likely to first appear.

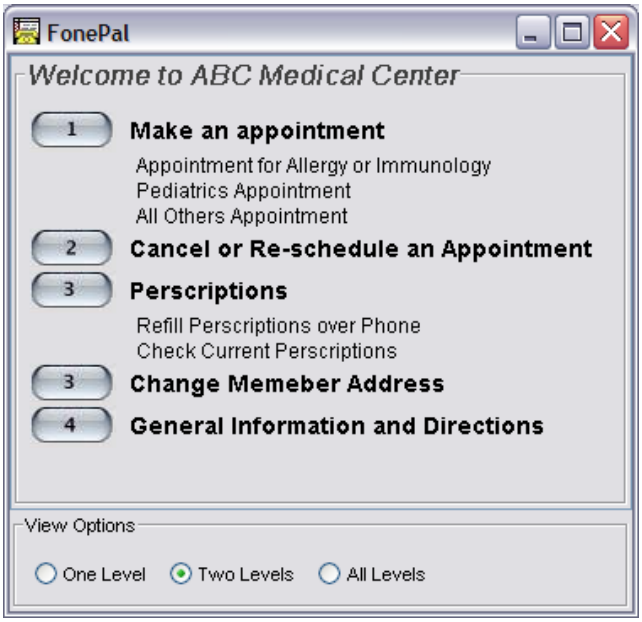

Figure 1: A screenshot of FonePal Client Window

Figure 1 shows an example screen shot of the client window which visually displays the text content of a voice menu. The buttons on the left hand side correspond to the choices the caller may currently select, which are also being spoken by the voice menu. Furthermore the current level menu items are emphasized with a larger bold font. Submenus are shown in a smaller plain font. If the caller selects one current menu item, the display will be updated and all submenu items of the selection will become the current menu items. The number of sublevels shown can be

varied in the "View Option" panel. One may find more system details of FonePal in Yin and Zhai [12].

Seated at a desk with a computer installed with a FonePal client, the participant was asked to call the simulated IBM helpline on a standard mobile phone connected to a loudspeaker. When the call was connected to our prototype IVR system through our local mobile carrier and standard PBX (Private Branch eXchange), the visual content of the voice menu was sent via AIM (AOL instant messenger) and displayed on the computer facing the participant. The participants could listen to the audio prompts and read the corresponding graphical menu. The visual menu always displayed the current and the next level content. The ability to display the next level submenus is one of the advantages of the visual modality. For voice menus, a more difficult tradeoff has to be made between the length of the voice message a caller can remember or tolerate and the indication of what the current choice really contain.

The visual menu had exactly the same content and structure as the voice menu, except for slight style changes from verbal to print expressions. Although the voice menus in IVR systems are usually designed with guidelines (e.g. [5, 9]) to accommodate the limitations of the voice modality, we decided not to change any of the menu structure even though a better visual design could be made for the same content, so that the usability of IVR systems in voice-alone mode would not be sacrificed.

The participants were asked to navigate the phone tree by pressing keys on the phone keypad, as a caller without visual menu does. In this and the next experiment, we did not enable the participants to use a mouse to interact with the visual menu on the computer screen so we could narrow our study to an experimentally manageable scope.

Methodology: The basic user study methodology we used in this experiment was "interaction history walk though". We let the experimental participants perform three trials of realistic phone tree navigation augmented by coordinated visual displays. These trials were recorded and later played back to the participants. During playback, the participants were asked to reflect and talk about what they did, how and why they did it, how they made their decisions, where their attention was, why they got stuck, if they were paying attention to the voice menu, and what improvement they would like to see. We preferred this method over simultaneous "think aloud" while performing the task because it does not interfere with their real time performance, particularly considering that our tasks involved voice output. To implement this method, we developed a "virtual VCR" that recorded all interaction events including all button presses on the phone and all voice and visual information presented during the experiment trials. Each trial of the experiment could be replayed with fast forward, backward, and pause functions.

Participants and task: six participants of different gender and job function ranging from research scientist to secretary were recruited from our lab. None of them had any prior experience with the experiment system. Their experience with the IBM helpline ranged from none to many times in the past. Their age varied from early 20's to mid 50's. They were given a three sentence explanation on what the system was with a sample screen shot of FonePal similar to Figure 1. We aimed to give the participants the amount of information a real user is likely to have before experiencing it – having heard about this new program and having just enough understanding of its functionality in order to install the software. To inform participants what is involved in using FonePal, they were shown the caller ID and IM ID registration step with default parameters before they proceeded to the calls.

The participants were given three written scenarios to call the simulated IBM helpline. In the scenario descriptions we tried to embed problems, such as setting up unified telephone/email messaging software, in the participants' minds as in real life. A trial was completed once the participant navigated the phone tree to a terminal node (a tree leaf) and was told "Please hold while your call is being transferred to a product analyst". The three scenarios were chosen by balancing various factors including 1. They were not the most frequently asked ones so the participants were unlikely to remember all of the correct choices from their previous experience. 2. The shortest path to the correct terminal node ranged from two to five key strokes so the three scenarios involved different levels of sub menus. 3. To accomplish all trials, the location of each correct choice at each level varied from the first (1) to the last (8) of the lists of available choices.

The participants were asked to navigate the IVR phone tree "as quickly as you can, just like in real 1-800 phone calls". After completing the three scenarios they filled out a short questionnaire. We then conducted the "interaction history walk through" with the participants, followed by an open-ended discussion and comments. These walk-through and interview sessions were voice recorded for later analysis.

Results

A rich body of qualitative data was collected in the walk-through and interview sessions. The following are some of the most informative highlights.

Initial experience: None of the participants had major difficulty relating the FonePal visual display to the voice menu due to the tight synchronization between the button presses on the phone, the voice menu prompt update, and the visual menu display refresh. The coupling between the phone and the computer display through IM was so tight that some participant could not even recall if they used the phone or the computer keyboard to navigate the phone tree. The participants quickly realized that the content on the screen was the same as what they heard on the phone. One participant commented: "*It took me a while to actually realize 'OK, I can just … go ahead and do what's on the screen*" (Participant S1-3). "A while" in this

case was still within the first minute of experiencing FonePal. Participants usually began to take advantage of the visual augmentation in the first few steps of the first phone call, and they obviously were ahead of the voice menu at least in some steps in the second trial.

However in the very first trial some of the participants did not seem to take advantage of the visual augmentation fully. This had two causes. First some were worried that the system wouldn't take their input if they did not wait for the voice prompt to finish. Second, some were not sure if the voice menu would have different or more information. These concerns disappeared in the second or third trial.

Specific advantages brought by visual augmentation: Some participants were articulate about the specific advantages with the visual menu. For example:

"What you often go wrong in these things is that there is a better choice down the road you didn't wait for it. But if you listen to them all then you …have so many it is hard to go about right? So you often take one if it sounds more or less right you take it first time through. Here [with visual menu] I can read them all …and pick the best one." (Participant S1-1).

"When you listen you have to remember everything ... because you think 'maybe there is a better choice' ... What's good about it [visual menu] is that you can see everything. you can see it all ...". (S1-2)

The participants clearly were taking advantage of the ability to scan back and forth. Commenting on how he made his decision on selecting Business Application for SAP:

"So I was like okay it is probably Business Application so I just scanned up and down quickly again just to make sure there is nothing else." (S1-3)

The role of the voice menu: With the visual menu displayed, the participants had equivocal opinions whether the simultaneous voice menu is an annoying disturbance or a help.

"I mean I wasn't listening. I was only reading to see … what the choice was. Once you have the idea that 'ah this is going to show you the choice on the screen', I didn't pay attention to what it was saying anymore." (S1-1)

"I don't care what she is saying right now [during second trial with FonePal]… I have to [ignore the voice] because she is literally going through choices that I know aren't even close to what are going to help me.… I mean they are drilling down and drilling down and drilling into some area it's just clearly not going be the appropriate choice " (S1-4).

"I didn't find it confusing, because ... I think once I decide to read something I tuned out what I heard" (S1-3).

On the other hand the same participant S1-3 commented later on a different episode where he first overlooked the option that contains the word "connectivity": "*My eyes were already down there* (near the end of the list), *then I heard connectivity, I am like oh and just pressed …"*(S1-3).

Similarly another participant did not see SAP on the visual menu at first. *"I missed SAP. … I didn't notice it until I actually heard it. … I scanned all the way down, didn't find SAP, then I start reasoning, … then I heard SAP, and somehow it prompt me to find SAP here [on the visual menu*]" (S1-5). This phenomenon of attending to the visual channel but still being alert to the voice channel upon hearing information of interest seems similar to the well known "cocktail party effect" whereby someone focused on a conversation can still hear his or her own name being mentioned in the noisy background.

Complaints about phone trees in general: A few of the participants also criticized and complained about the content and classification of the helpline used in our experiment, which were copied from a real system. Designing IVR phone trees is a difficult challenge – the customers want them concise, logical, and descriptive in their own words which may differ from one person to another and to the designer. For example for a networking problem some in the study looked for help with the term "network" or even "Web" in their mind while the IVR menu used the term "connectivity". Our participants also pointed out that problems do not always fall logically into only one branch of the IVR menu tree:

"Now (referring to his hesitation at one point) this is a problem of course always. Is it workstation support? Is it business application?" (S1-1)

Overall Reaction to visual augmentation: Overall, the participants were enthusiastic about the benefits of visually augmenting voice menus. They unanimously answered yes to the question "If you find FonePal helpful". On the scale of 1 (not at all) to 5 (very helpful), they unanimously selected 5. Individually, they made the following comments:

"I would love to have it!" (S1-5).

"The visual does help because … I mean it's a fantastic feedback. … because you don't have that just audio-wise. " (S1-4).

"I think this will be quite useful" (S1-3).

"*This is definitely helpful. … It was pretty obvious what to do. This is very helpful"* (S1-2).

"*I would use this, a lot. The phone makes me crazy. It is irritating that the phone takes up so much of my time. With this I could look at it and 'Ok got it. Ok got it'"* (S1-6).

EXPERIMENT 2

Complementing Experiment 1, this experiment was focused on the quantitative performance of the visual augmentation of voice menus in comparison to voice menu alone.

Tasks, conditions, and experiment design

In addition to the IVR menu (IBM-HELP) and scenarios used in Experiment 1, another menu tree, copied from the IBM external customer service line 1-800-IBM-SERV, was added in this experiment, along with three new scenarios comparable with the first three scenarios in complexity. The participants were less likely to be familiar with the IBM-SERV menu, but the scenarios chosen were related to common PC software and hardware problems, such as having purchased a hard drive that may need replacement.

Sixteen participants of the same demographic and job variation were recruited for Experiment 2. None of them participated in the first experiment. This two-condition within-subject experiment was balanced with regard to condition order (With or without visual display first) and phone tree (IBM-HELP vs. IBM-SERV). Each participant was asked to read six scenarios and make six phone calls, three to IBM-HELP and three to IBM-SERV, three with visual augmentation and three without. If the three phone calls a participant made to IBM-HELP were those with visual augmentation, then the next three to IBM-SERV were without visual augmentation. For the next participant, the calls to IBM-HELP would be without visual augmentation and the ones to IBM-SERV with visual augmentation. Altogether 96 phone calls were made, of which 48 were with visual (24 to IBM-HELP and 24 to IBM-SERV), and 48 without visual augmentation (24 to IBM-HELP and 24 to IBM-SERV).

Results

Completion time and error rate were two basic performance measures used. The completion time was counted from the moment the call was connected to the moment the last key was pressed (DTMF signal received) before the call was transferred or disconnected. A trial was considered an error trial if the destination could not reasonably match the scenario given.

Completion time

Repeated measure variance analysis shows that the mean task completion time changed significantly with the experimental Condition: $F_{1,14} = 24.6$, $p = .0002$. With the visual menu (FonePal), the mean completion time was 36% shorter than without FonePal (Figure 2). The impact of the Order of the condition tested on completion time was not significant ($F_{1,14} = .823$, $p = .38$), neither was the impact of the order's interaction with Condition ($F_{1,14} = 2.59$, $p = .13$).

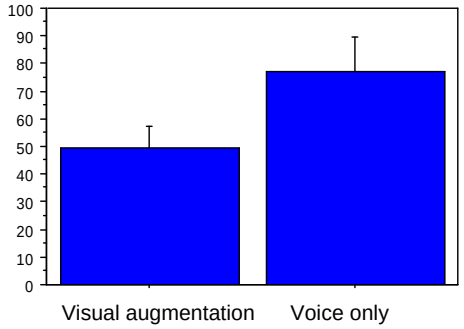

Figure 2: Mean and 95% confidence interval of completion time (in sec) with and without visual menu.

Error rate

We first analyzed the error rate at a very coarse level: a trial (of going through an entire scenario) was either considered completely correct or completely wrong. At this level, the participants failed 12 out of 48 trials without visual augmentation (25%) and 3 out of 48 trials with visual augmentation (6.3%), amounting to 75% reduction in failed trials. This difference is statistically significant by the non-parametric Fisher's exact test ($p = .022$).

We then analyzed error rate at a finer, partial error rate level. Instead of judging a trial as a complete failure or complete success, we scored each trial based on the number of correct steps accomplished and the logical proximity the terminal node selected was to the correct answer. Figure 3 shows that with visual augmentation the mean (and 95% confidence interval error bar) of the partial error rate decreased from 12.46% to 3.21%, amounting to 74.24% reduction in partial error. Tested by repeated measure variance analysis, this difference is statistically significant: $F_{1,14} = 5.1$, $p = .04$. Neither the Order of the condition tested ($F_{1,14} = .15$, $p = .71$) nor its interaction with Condition ($F_{1,14} = .39$, $p = .55$) was significant.

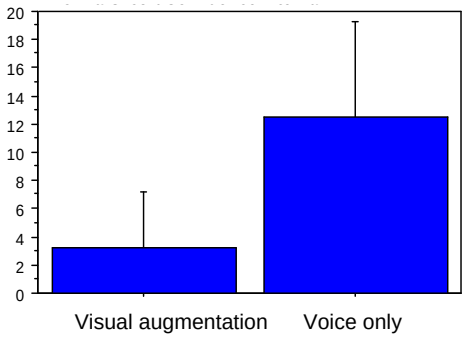

Figure 3: Partial error rate (in percentage): Mean and 95% confidence error bar

In depth analysis – comparison to normative voice time

We observed that with visual augmentation the participants would frequently make a correct selection well before the voice reached or finished speaking that selection. We divided all correctly completed trials into multiple steps corresponding to the multiple levels of the menu tree, and defined a "normative time" of each step from the moment a menu is presented to the time the correct choice in the menu has been spoken and measured each individual's actual time spent on that step. The quntitative relation between the actual and the normative time could inform us whether and how much the callers took advantage of the visual augmentation. Note that with voice alone one could also beat the normative time of a step if the caller memorized the choices from the past. For example most participants did not wait for the completion of "Please enter your six digit IBM serial number …" before they started entering the employee number given to them, with or without visual augmentation. On the other hand, one could be much slower than the normative time even with visual augmentation when deciding the best among several reasonable choices.

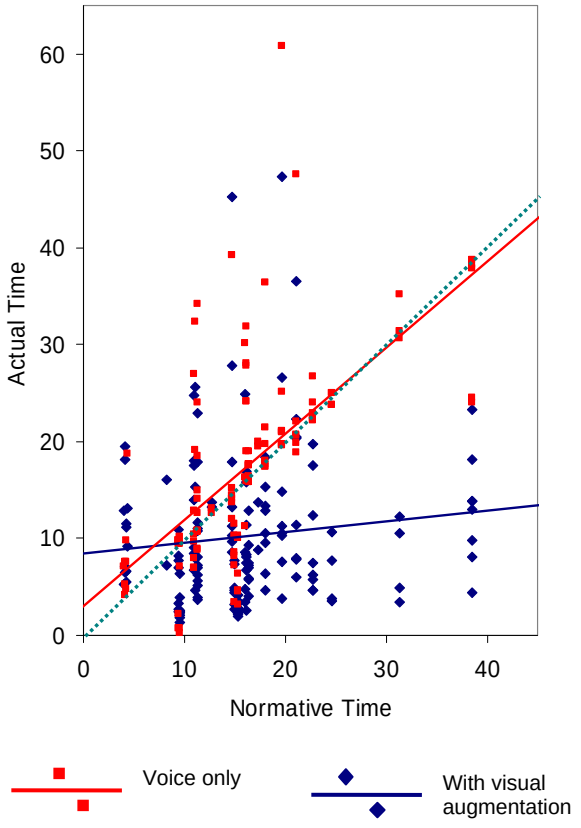

Figure 4: Scatter plot of actual against the normative time (sec) of all navigation steps taken in all trials. The red upper solid line is linear regression for voice only, and the blue bottom line is with visual augmentation. Dotted is the reference line y = x.

Figure 4 shows the scatter plot of actual time vs. normative time in all correct steps taken by the 16 participants in Experiment 2. The linear regressions between the actual and normative times are:

Voice Only:

$$t_a = 2.98 + 0.89\ t_n \text{ (sec)}, \quad R^2 = 0.42; \qquad (1)$$

With Visual Augmentation:

$$t_a = 8.46 + 0.11\ t_n \text{ (sec)}, \quad R^2 = 0.013; \qquad (2)$$

where t_a is the actual time taken in each step and t_n the normative time for that step.

With voice menu only, the normative time accounted for a large portion of the variance in the actual step time (42%). It is quite plausible that the normative time accounted for some but not all of the variance in the actual time, since in most cases callers wait until the appropriate choice is announced before making a selection. On the other hand, callers could either remember the choices from previous trials and select faster than the normative time or wait all the choices are announced then ponder them before making a decision hence act much slower than the normative time.

In contrast, with visual augmentation, the normative time accounted for very little of the variance in the actual selection (1.3%). This suggests that the participants did not pay much attention to the voice menu and they could scan the visual menu so fast that the relative location of the appropriate choice on the list mattered very little.

Each step in navigating the phone tree took from a few seconds to about a minute, but visual augmentation was clearly a major determinant. Only 14.3% of the steps taken without visual augmentation were "clearly" (by at least 5 seconds) faster than the normative time. Most of these happened at the step "Please enter your six-digit IBM serial number" and the step "You've entered 1-2-3-7-8-9. Press 1 if this is correct. Otherwise press star followed by your six-digit IBM serial number". In contrast, fully 60% of the steps in the FonePal condition (with visual augmentation) clearly beat the normative time.

As shown in Figure 4, linear regression also indicates the trend that the longer the voice menu is (longer list / longer average normative time), the more time savings the visual augmentation would offer.

Subjective rating and open comments

On the scale of 0 (not at all) to 5 (all the time), 12 of the16 participants (75%) said they would like to have visual augmentation all the time (5), 2 (12.5%) said they would like to use visual augmentation frequently (4), 2 (12.5%) said they would like to use visual augmentation often (3). Participants of Experiment 2 were also asked how they handled both the audio and visual channel when working with FonePal. Of the 16 participants, 4 (25%) said they based their selections on what they "see on the screen only"; 10 (62.5%) said they based their selections "mostly on what I see, but occasionally also what I hear"; 2 (12.5%)

declared to have made their selections based on "both what I hear and what I see". When asked whether the voice was distracting when using visual augmentation, 3 (18.75%) felt "somewhat", 8 (50%) felt "a little", and 5 (31.25%) answered "not at all".

In this experiment we also conducted interviews with all participants and did "interaction history walk through" with selected episodes either with or without visual augmentation that the participants wanted to talk about. Their comments largely concurred with those in Experiment 1 and were clearly in favor of visual augmentation of voice menus. For example, one participant commented: "*It took me two seconds [to decide] I wanted to have it. It is just so clearly better*". Interestingly this participant completed all the voice only scenarios successfully and rather quickly but he still preferred visually augmented voice menus. "*When I call [with voice menu only] I found it stressful. If you had a blood pressure meter on you might have seen it when it was in fact very stressful… I mean I so hate to listen to a long serial thing…. that I'll make a choice and it will be awful …I will hang up and call again*". Speaking on the frustration he has with navigating IVR systems and the benefit of visually augmented voice menus, he further commented somewhat humorously that "*This might be one of the most important pieces of work you ever do because people might live longer … it is such a good project. I mean there is no doubt I would want to have it … there is no down side. It is a complete plus!*"

In summary, with the two phone trees tested which mirrored real corporate help lines, on average the participants were 36% faster in navigating phone trees with about 75% fewer errors. With visual augmentation the participants made their selections clearly ahead of the voice menu at least 60% of the time. Due to the advantage of fast and self-paced visual scanning, with visual augmentation the relative location of the correct choice on a list mattered little whereas with voice menu only this was a major determinant to the completion time. The subjective ratings and comments concurred with these conclusions.

EXPERIMENT 3

A visual channel affords more possibilities than menu selection by browsing. Keyword search is another alternative means to find the terminal node (leaves) of the phone tree. Indeed, during our first experiment, a few participants commented that they would like to have a search function to directly retrieve the desired terminal node instead of browsing the voice menu level by level. Browsing vs. search has long been a topic of interest in HCI research, particularly in the early days of the web [4]. The effect of search obviously depends on the size of the information space. For the immense amount of information on the web, users quickly turned to search as the primary means of finding information. More recently, web style desktop search is also emerging. In contrast activating commands and applications on a PC remains almost exclusively the action of visual manual browsing (which could change as the number of software applications is ever increasing). We are interested in whether search will be a viable choice in the context of navigating phone menu trees with a moderate to large of number of choices. Will search be more helpful or more efficient than browsing in this context? Will people prefer one over another? To answer these questions, we conducted the third experiment.

Tasks and Conditions

Experimental System: For Experiment 3, we implemented a search function in addition to the existing browsing function in FonePal. Figure 5 shows the new client window with the added search function.

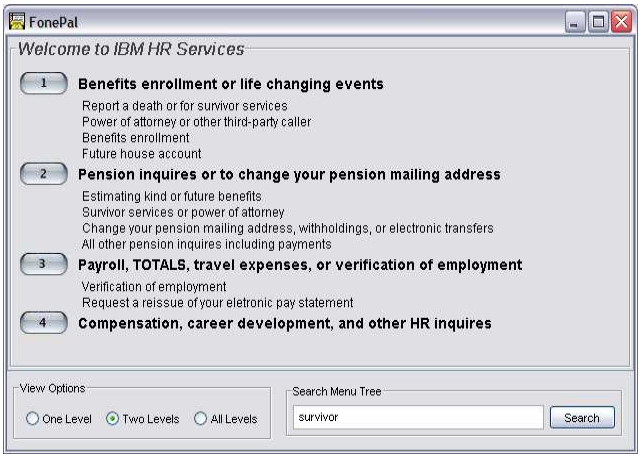

Figure 5. Client Window with Search Function

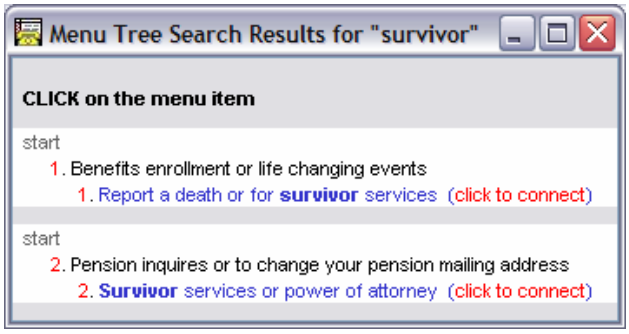

Figure 6. Search Result Window

If a caller wants to skip the middle levels of the voice menu tree and jump directly to a terminal node, s/he can type one or multiple keywords into the "Search Menu Tree" box (Figure 5). The search result is shown in a separate window. For example, if a caller phones "Human Resource Benefits Inquiry" and types the keyword "survivor", two relevant terminal nodes of the phone tree together with their paths will be displayed to the caller, as shown in Figure 6. The

caller can then click on one of these nodes to be connected directly to the appropriate representative.

In our prototype system a simple and basic keyword matching schema was used to search the phone tree for the appropriate terminal nodes along with the paths from the caller's current position to the terminal nodes. The keyword matching was based on the keywords entered by the caller and the text contained in the phone tree. Multiple search results were sorted according to their rank. Paths that contained more keywords ranked higher than those with fewer matched keywords. Paths also ranked higher if the matched keywords were contained in the text of their terminal node. In addition, matches of different keywords were not treated equally. Words that commonly appear in the entire phone menu (e.g. "of") were considered to carry less information and therefore weighted less when computing scores for ranking.

In order to increase the search quality potentially more context or descriptive information could be added to the phone tree that are accessible only to the search function. But the quality and quantity of such information has to be carefully controlled. While finely crafted supplemental information can be helpful, in this study we were interested in the general effect of searching a typical phone tree that is not custom engineered.

Method: The experimental task was to place phone calls and navigate IVR phone trees with FonePal browsing and FonePal search respectively. FonePal browsing was the same as in Experiment 2. For FonePal search, participants were asked to form one or more keywords by themselves based on the given scenario. A trial was completed by clicking on one of the returned search results. Participants could try different keywords if the returned results did not contain what they were looking for.

Each participant practiced using browsing and search based on a real phone tree, the IBM Human Resources helpline. This phone tree was used for the practice trial only to prevent possible knowledge transfer of the menu structure and content. For data collection, we used the same voice menus and task scenarios as those in Experiment 2: three scenarios with 1-888-IBM-HELP and three scenarios with 1-800-IBM-SERV. In each scenario, the participant was presented with a computer-related problem, for example, needing to replace a noisy hard drive. The participant was asked to place a phone call to our simulated IVR system then browse or search the menu tree to find the right terminal node. The experiment was balanced between browsing and search. If participant A was asked to use browsing with the three scenarios of 1-888-IBM-HELP and search with the three scenarios of 1-800-IBM-SERV, the next participant, B, would be asked to use search with 1-800-IBM-HELP, and browsing with 1-888-IBM-SERV.

Participants

Sixteen participants of both genders and varying job functions ranging from research scientist to human resource specialist were recruited from our lab. None of them had prior experience with FonePal or participated in the first two experiments. Their demographics were similar to the participants in Experiment 2. At the end of the study, the participants were asked to fill in a questionnaire. They were asked if they felt browsing or search were helpful and their preference among "Without FonePal", "FonePal Browsing", and "FonePal Search".

Results

In addition to the questionnaire, completion time and error rate were used as quantitative measurements to compare visual manual browsing and keyword search.

Completion time

Although on average search was faster than browsing (Figure 7) on the tasks and phone menus tested, repeated measure variance analysis shows that the difference between these two conditions was not statistically significant: $F_{1,15} = 2.21$, $p = .16$. On average the completion time was slightly faster than the FonePal trials in the second experiment, probably a contribution of the additional practice trials unique to Experiment 3.

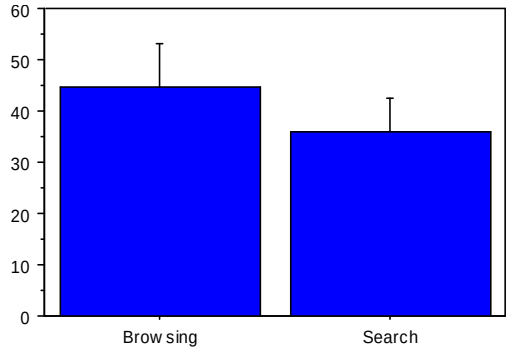

Figure 7: Mean and 95% confidence interval of task completion time (in second) with browsing and search

Error rate

Since it is difficult to divide a search trial into individual steps, we did not calculate and compare partial error rate as we did in Experiment 2. If we consider a trial as either completely right or completely wrong, the error rates of the two conditions were exactly the same, both at 8.33% (4 out of 48 trials). This is also about the same as the error rate in Experiment 2.

Subjective evaluation

Experiment participants also had mixed views on the relative merits of each method. All 16 participants found FonePal browsing helpful and 15 of the 16 participants found FonePal search helpful. Overall they have the same positive response toward visual menu augmentation. *"I*

think the function is really good, I really like it." "*… with FonePal is definitely better than without FonePal.*" "*It saves time, a lot of time.*" In search trials, on average it took 1.56 attempts to find a desired terminal node, as some participants tried out different keywords when the first attempt was unsuccessful. When asked their preference, 7 of the 16 participants preferred search. 8 of the 16 preferred browsing. Participants who preferred browsing seemed to consider the certainty of viewing all possible choices in context as more important:

"[With browsing,] I know everything is here, and I'm very confident."

"'cause if you don't know what you are looking for, browsing kind of gives you, … a lot more hints than hope you get the right search criteria the first time."

"Because it [browsing] is prompting, instead of me having to search for it."

"People don't like to type something, just want to click."

It is worth mentioning that there were two participants who were successful with all of their first search attempts, but they still preferred browsing to search. This suggests that the participants may still prefer browsing even though search is at its best performance, and there might be other key factors that worth looking into. For those who preferred search, they seemed to be confident that FonePal would understand their keywords (eventually):

"I like search because it is fast, you don't have to read [browse] all the menus. It takes you to a few choices."

"I like it [trying out different keywords] in search."

"It [search] is much better, actually. You get more information up there, … I felt I had closer choices than I did if I punch their menu."

One participant preferred "*search first then browse*". Others were also interested in combining the two but would take different strategies. "*But if they give me the prompt, it may not have what I need, but then there should be a function in there that says "None of the above", so then you go to your search and you can input.*" "*…,in time, I'd probably use search often too.*"

In sum, the average completion time and error rate did not differ significantly between searching and browsing for the two phone trees and task scenarios tested. However the experiment suggests that keyword search and visual-manual browsing have different but complimentary benefits. Keyword search is likely to be faster if the menu is more complex but it tends to introduce the uncertainty of hit and miss. In contrast, visual manual browsing is more assuring to some users since they can be certain of what they had navigated through. Since the two methods are not technically mutually excusive, both should be available to users of FonePal types of applications and systems.

DISCUSSION AND CONCLUSIONS

Navigating IVR phone trees in order to reach the human agent is a common interaction experience. Often it is difficult and frustrating due to the sequential and fixed paced nature of voice menus. The three experiments presented here clearly demonstrate that this common interaction experience can be significantly improved by augmenting telephone voice menus with coordinated visual displays. As shown in Experiment 1 callers could be much more satisfied with and even enthusiastic about their augmented voice menu navigation experience. As shown in Experiment 2 the navigation time savings and error reduction can be quantitatively measured and modeled as a function of normative time. The more complex the menu tree, the greater the advantage of visual augmentation. With visual augmentation keyword search can also be easily enabled, which complements visual-manual browsing as demonstrated in Experiment 3. Although the results of Experiment 1 and 2 may seem to be unsurprising to some, we consider these experiments necessary and the results valuable. This is because human performance can be counter intuitive. Without experiments we cannot tell if there is actually a performance gain associated with the visual augmentation and more importantly how much the gain will be for typical business phone trees. The data set and regression model obtained from Experiment 2 result may also be used to make an informed decision on whether or not it is worthwhile to engineer a visual augmentation solution over the existing phone tree.

We expect the benefits of augmenting phone calls with a visual channel to be greater in real world tasks. In laboratory studies, participants were more focused on navigating the phone trees than in real life. Interruptions and multi-tasking (e.g. listening to an IVR menu and glancing at email) in real world situations can make navigating a voice menu even harder since critical information in the voice menu can come up when the user's attention is focused elsewhere.

The technical and system feasibility of instrumenting IVR systems with FonePal, particularly by means of cross-device integration orchestrated over the Internet, has been previously discussed [12]. Mobile or IP phones with large screens are increasingly common. Without the problem of device and ID association, visual augmentation of voice calls on these phones is also possible and to some extent easier to implement.

We have limited FonePal functions to IVR menu navigation so the studies were within an experimentally manageable scope. Obviously once a visual channel is automatically set for the caller, more multimodal and multi-channel functions can be supported. Some of our experiment participants suggested embedding hyperlinks on the FonePal panel to the current "Most frequently asked questions" which can be customized according that caller's profile and history. The search function can also be easily extended to online help materials beyond the phone tree. In some cases, the caller

may be able to find enough useful information before speaking to a human agent, saving time and cost for both the customer and the service provider. With careful design, it is also possible to carry over the multi-channel set-up after being transferred to a human agent.

In summary, augmenting phone tree navigation is not only technically feasible as demonstrated in [12], but also clearly beneficial from a human performance and user experience perspective as demonstrated in the series of studies presented here. These benefits could potentially translate to large reductions in the cost of handling incorrectly routed calls and greater satisfaction to millions of callers.

ACKNOWLEDGMENTS

The authors would like to acknowledge the contributions of John Barton, Steve Cousins, John Day, Per-Ola Kristensson, Dan Shiffman, Barton Smith, Alison Sue, Pernilla Qvarfordt and the participants of our studies.

REFERENCES

1. Fawcett, P.E., Blomfield-Brown, C. and Strom, C.P. System and method for graphically displaying and navigating through an interactive voice response menu, US Patent 5802526, USA, 1998, 1998-1909-1901.
2. Hillier, C. Text-enhanced voice menu system, US Patent 6493428, USA, 2002.
3. Kreitzer, S.S. Voice Augmented Menu Automated Telephone Response System. *IBM Technical Disclosure Bulletin*, *38* (2). 1995,57-62.
4. Mackinlay, J.D., Zellweger, P.T., Chignell, M., Furnas, G. and Salton, G., Browsing vs. search: can we find a synergy? (panel). *Proc. ACM CHI'95 Conference on Human Factors in Computing Systems*, *1995*, 179-180.
5. Marics, M.A. and Engelbeck, G. Designing voice menu applications for telephones. in M. Helander, T.L., P. Prabhu ed. *Handbook of Human-Computer Interaction*, Elsevier, Amsterdam, 1997, 1085-1102.
6. Narayanaswami, C. Graphical voice response system and method therefor, US Patent 6104790, USA, 2000.
7. Paap, K.R. and Cooke, N.J. Designing menus. in M. Helander, T.L., P. Prabhu ed. *Handbook of Human-Computer Interaction*, Elsevier, Amsterdam, 1997, 533-572.
8. Resnick , P. and Virzi, R.A., Skip and scan: cleaning up telephone interface. *Proc. ACM Conference on Human Factors in Computing Systems*, *1992*, ACM, 419-426.
9. Roberts , T.L. and Engelbeck, The effects of device technology on the usability of advanced telephone functions. *Proc. ACM CHI conference on Human factors in computing systems*, *1989*, 331-337.
10. Suhm, B., Freeman, B. and Getty, D., Curing the menu blues in touch-tone voice interfaces. *Proc. Extended Abstracts of ACM CHI Conference on Human Factors in Computing Systems*, *2001*, 132-133.
11. Whittaker, S., Hirschberg, J., Amento, B., Stark, L., Bacchiani, M., Isenhour, P., Stead, L., Zamchick, G. and Rosenberg, A., SCANMail: a voicemail interface that makes speech browsable, readable and searchable. *Proc. ACM CHI conference on Human factors in computing systems*, *2002*, 275-282.
12. Yin, M. and Zhai, S., Dial and see: tackling the voice menu navigation problem with cross-device user experience integration (TechNote). *Proc. UIST 2005 -- 18th ACM Symposium on User Interface Software and Technology*, *2005*.