The effectiveness of human and synthetic hybrid Voice User Interface (VUI) Prompts

Thomas G. Petersen
University of Missouri - Rolla
tgpdmb@umr.edu

Richard H. Hall
University of Missouri - Rolla
rhall@umr.edu

Abstract

The goals of this project were to: a) determine if changing the voices speaking from a recorded human voice speaking information to Text-To-Speech (TTS) for navigation provided an audible equivalent to a visual hypertext link as an available option and b) evaluate the effect on age groups and genders. Participants dialed a toll-free number and entered or spoke the provided codes. Either a hybrid or synthetic interface type was assigned to the phone number. After four stimulus questions, the voice activated cheap gas price finder activity became active for the phone number where the participants were asked to provide a 5 digit zip code of their choice. Callers had the ability to sequence through the results, filter by town and hear the address of the station. The results did not identify a significant difference between interfaces. An unanticipated outcome where some participants spoke longer, more complete responses will be evaluated in a future study.

Keywords

Interactive Voice Response (IVR), VoiceXML, Text-to-Speech (TTS), Voice User Interface (VUI)

Introduction

Background

In their book Wired for Speech, Nass and Brave (2005 p. 3) claim we have become "voice activated" and that "talking, listening, and human society have elegantly coevolved into a remarkably interwoven, effective, and stable system." With the innate propensity for voice communication between humans, the next logical step for dialog between a human and a machine is the aural version of the visual Graphical User Interface (GUI) called the Voice User Interface (VUI) in the computer speech industry. To facilitate a request/response spoken dialog between human and machine, VUI technology prompts the user with Text-To-Speech (TTS) or a pre-recorded narration then waits for a voice response from the user. Upon recognizing a voice input, the machine interprets the utterance, matches it with the associated action and the cycle repeats.

Due to the hands-free nature of speech, VUI driven devices lend themselves to applications where the input device does not provide adequate physical input functionality or the user is unable to provide input through the device interface due to proximity, distraction, or lack of human physical or visual ability. For example people, with limited vision or wearing protective equipment that restricts vision or mobility or that complicates manual data entry.

Rationale

Like a human-to-human conversation, dialog with a VUI is linear with facts, available choices and facts about the available choices (and inferred facts from previous experience) presented during the discourse. Included within the information body presented during oral communication are variances in pitch, word choice, and inflection to signify statement type such as a command or a question. Oftentimes one person may speak specific "hot words" as offerings that the other person may or may not retrieve and implement by repeating the same hot word(s) or synonym(s) to confirm acceptance/agreement. One of the biggest challenges with effective VUI design is communicating to the user what hot word(s) the machine expects to hear without being redundant. As Donald Norman points out in his book, The Design of Everyday Things (2002), interaction with the device should be obvious by its design. Several studies have been performed with mixed initiative systems (hybrid) where human voice was used for static content and TTS delivered dynamic content but little or no research evaluates a hybrid system with the machine speaking the hot words and the impact on usability for age groups and gender. In the aural channel of communication where sound is the only means to convey navigation instruction and information simultaneously, mixing human voice for instruction and synthetic voice for navigation solves the redundancy issue without reducing functionality.

A study evaluating user trust, likes and competency of a call-in housing information system using all TTS delivery and a hybrid system with a human voice speaking static content and TTS speaking the dynamic information revealed the all TTS system was trusted significantly more then the hybrid and the TTS system sustained a significant preference for liking and competency. When the recorded voice system was added, it rated significantly higher than the TTS as likable, trustworthy and competent (Nass, et al., 2005, p. 254).

When combining TTS audio and human visuals, both genders found the consistent combinations more trustworthy (Nass, et al., 2005, p. 246). "[…] woman felt that the inconsistent talking head was more upsetting and more strange and rude relative to the consistent talking head then men." In a different study by Nass et al. (2005), voice gender of a voice only system influenced the user's perception of trustworthiness and conformity that were "positively oriented toward synthetic voices whose 'gender' matched their own" (Nass, et al., 2005, p. 15).

In addition to research of synthetic audio visual combinations Nass and Brave have a chapter dedicated to combining human and TTS "Mixing Synthetic and Recorded Voices: When 'Better' is Worse" (p. 143) yet do not discuss the combination for navigation in their book "Wired for Speech" (Nass et al . 2005). The housing information study results indicate consistency, specifically, human voice, is preferred when delivering information but nothing mentioned about the design of navigation in content delivery.

Studies performed on various design factors of VUI highlight aspects that work well and identify drawbacks. The chosen type of VUI design is highly task dependant addressing only the time based aural sensory channel. Therefore, the system must communicate a clear interaction model (Tomko et al. 2005), navigation structure (Bolchini et al. 2006), and choice feedback including current progress/location using only the aural sensory channel (Perugini et al. 2006; Walker, et al. 1998).

Approach

A plethora of techniques exists when communicating visually such as underling text or various state changes signify something is clickable on a web page; however defining the auditory equivalent to a hyperlink poses a few challenges. Previous research points out the importance of communicating the available options and what users must say access them, but a common strategy or standard similar to the visual indicators of the hyper-link does not exist for VUI's.

This paper will explore an implementation of Norman's philosophy that function should be obvious from design with two types of voice interface design:

1. Male human voice for static content and female TTS for dynamic content and the expected responses spoken with male TTS.
2. Male TTS voice for static content and female TTS for dynamic content and the expected responses spoken with the same male TTS.

- What is the effect of voice interface type on performance?
- What is the effect of voice interface type on call duration?
- What is the effect of voice interface type on VUI satisfaction?
- Does gender have an impact on voice interface usage?
- Does age have an impact on voice interface usage?

Method

Participants

At the macro level, anyone living in the continental United States of driving age and cell phone user/owner meets the participation requirements. A research participation invitation was sent via e-mail to distant friends, co-workers, classmates, colleagues and family. Research participation requests also included users of the goog411 and gatewayCHI user groups and fifty flyers handed out at the Saint Louis Science Center during World Usability Day 2007. Mid-west residents had the highest response rate but participation included both east and west coast inhabitants.

Respondents:

	Male	**Female**	**Total**
18-29	5	4	9
30-49	5	5	10
50-54	3	4	7
Total	13	13	26

Materials

The BeVocal Café is a VoiceXML development platform providing a free, Web-based environment for testing and development. For each call, BeVocal provides a recording of the call and a detailed log accessible through a web interface.

Four stimulus questions collected simple demographic data about the participant. Response times and accuracy measured performance with the two interface types:

- Text-to-Speech (TTS)
- Hybrid (recorded human voice for content and TTS for available responses)

.

I am in the age bracket:

- 18-29
- 30-49
- 50-64
- or over 65

I am:

- male
- female

When using my cell phone and driving I:

- Never use it
- Hold it
- use an Ear piece
- use a headset
- or use the Speaker.

I buy gas based on:

- convenience
- price
- brand

Procedure

Participants received a message via e-mail, reading a posting at goog411 or gatewayCHI user groups or received a flyer at the Saint Louis Science Center on World Usability Day 2007. The message invited people to participate in a student research project for the University of Missouri – Rolla by calling a toll-free number (BeVocal) and providing the codes when prompted.

- *DIAL:*
  (877) 338-6225
- *SPEAK:*
  USERID: 1516444
  PIN: 1516

Once granted access to the system, one of two VUI types was assigned to their phone number. A recorded human voice provided a brief overview of the research purpose and read the URL of a text version of the consent form that was read to them. Speaking their name when prompted signified verbal agreement to the terms and conditions, the utterance was recorded with the session call logs.

An announcement, "the survey will now begin" by a male TTS voice served as a transition to the survey. A female voice provided a status update before each question. Depending on the assigned interface type, hybrid, or TTS, participants either heard the question spoken by a recorded human voice and the available responses read with TTS or the TTS version which used TTS for the question and responses.

Upon survey completion, participants were informed they will move on to the next part of research, the cheap gas price finder and subsequent calls will bypass the survey. While in the cheap gas price finder, participants were prompted for a five digit zip code they wanted to find the cheapest gas price. If the assigned VUI was hybrid, a recorded human voice spoke the static information, and a male TTS voice the expected responses, otherwise it was all TTS.

A female TTS read the total number of stations found, the cheapest price, location, and last update. If the assigned VUI was hybrid, a recorded human voice spoke the static navigation information, and a male TTS voice spoke the expected responses, otherwise it was all TTS. If more than one station was found the available options included:

- **next**
  (female synthetic spoke price, brand and price update date for the next station)
- **previous**
  (female synthetic spoke price, brand and price update date for the previous station)
- **first**
  (female synthetic spoke price, brand and price update date for the first and least expensive station)
- **last**
  (female synthetic spoke price, brand and price update date for the last and most expensive station)
- **address**
  (female synthetic spoke street address for the current station)
- **filter by town**
  (male synthetic spoke towns with available station data in the zip code, users spoke the town)
- **zip code**
  (prompted for five digit zip code to obtain cheapest gas price)
- **new**
  same as zip code
- **options** (available at all times)
  (male synthetic spoke words available at the current prompt)
- **help** (available at all times)
  (all details of application navigation with interface voice)

Sessions terminated when callers hung up.

Results

Although participants answered specific questions, only responses to age and gender questions were measured. To examine the effects of age and gender on Voice User Interface (VUI) type (hybrid vs. TTS), duration, and errors, a series of T-tests were performed on the data with total errors and duration serving as the dependant variables and interface, gender and age group as the independent variables. A significant difference for the interface type as hoped. Satisfaction and VUI preference was measured by repeat calls, which was zero.

		18 to 29			
		TTS		HYBRID	
		Male	***Female***	***Male***	***Female***
Age:	***Error Count:***	0	1	0	0
	Mean Time:	.693	.500	.000	1.829
Gender:	***Error Count:***	0	0	0	0
	Mean Time:	.662	1.774	1.156	.938
Driving:	***Error Count:***	0	1	0	0
	Mean Time:	2.490	1.211	.243	.766
Purchase Habit:	***Error Count:***	1	0	0	0
	Mean Time:	1.656	2.554	1.469	2.001
Zip:	***Error Count:***	0	0	1	0
	Mean Time:	2.037	2.211	2.344	2.883
	Mean Error Count:	.2	.4	.2	0
	Mean Time:	1.507	1.65	1.042	1.683
	Participants:	3	2	2	2

		30 to 49 TTS		HYBRID	
		Male	***Female***	***Male***	***Female***
Age:	***Error Count:***	0	0	0	0
	Mean Time:	.758	1.914	1.870	1.281
Gender:	***Error Count:***	1	0	0	0
	Mean Time:	1.516	1.391	1.672	.969
Driving:	***Error Count:***	2	1	1	0
	Mean Time:	1.477	.852	1.417	1.646
Purchase Habit:	***Error Count:***	0	0	0	0
	Mean Time:	2.163	1.469	1.776	1.932
Zip:	***Error Count:***	0	0	0	0
	Mean Time:	1.999	2.157	1.802	1.980
	Mean Error Count:	.6	.2	.2	0
	Mean Time:	1.583	1.556	1.707	1.561
	Participants:	2	2	3	3

		50 to 64 TTS		HYBRID	
		Male	***Female***	***Male***	***Female***
Age:	***Error Count:***	0	1	2	0
	Mean Time:	0	1.261	2.068	1.125
Gender:	***Error Count:***	0	0	0	0
	Mean Time:	0	1.234	1.818	.953
Driving:	***Error Count:***	3	0	0	2
	Mean Time:	0	1.245	2.239	1.594
Purchase Habit:	***Error Count:***	0	0	0	0
	Mean Time:	0	1.995	3.032	2.969

Zip:	***Error Count:***	0	0	0	0
	Mean Time:	0	1.562	1.844	1.750
	Mean Error Count:	0	.2	.4	0
	Mean Time:	0	1.459	2.200	1.678
	Participants:	0	3	3	1

	ALL TTS		**HYBRID**	
	Male	***Female***	***Male***	***Female***
Total Error Count:	7	4	4	2
Mean Time:	1.1325	1.555	1.650	1.640
Participants:	5	7	8	6

Discussion and Conclusions

Although the results of these tests did not indicate a significant difference between the two interface types relative to performance or call duration, as with some of the Nass work, much was learned. The required seven digit *user id* and four digit *pin* were an Achilles heel in many aspects. Participants were required to memorize or retrieve the numbers for each call making the system inconvenient to use right-off-the-bat. Satisfaction and VUI type preference was to be measured by return calls, which was zero and attributed to the access codes. The order in which the codes were requested changed from call-to-call with a likelihood of frustrating users forcing users to pay attention to a mundane task. Some people entered the codes with the dial pad and BeVocal would audibly remind the users to speak the *user id* or *pin* after each key that was pressed, but would accept dial pad entry anyway. With the phone was away from the users' ear to enter the codes, did the participants even hear the reminder? The BeVocal entry gateway seemed to have difficulty understanding some people, specifically participants whose native language is not English. It is quite possible the BeVocal gatekeeper turned many potential participants away, native English speakers and all. Some people experiencing recognition conflicts resorted to entering the codes with the dial pad despite reminders to speak the codes.

People have become accustomed to TTS and the qualities of synthetic voices are almost indistinguishable to human voices, it is quite possible participants did not notice the difference. Future tests could be performed with opposing voice genders in a hybrid interface to further differentiate information words from navigation words.

When evaluating, the call data, some participants actually spoke the expected response in its entirety, some partially. For an example "filter by town" was provided as the choice, some people said, "town," "filter" or "filter by town," all three were accepted. If there is a commonality, this phenomenon reinforces a portion of the hypothesis that separating voice types for content and navigation does have an impact when interacting with a VUI.

Acknowledgments

This research was supported in part by the BeVocal Café.

References

Nass, Clifford, Brave, Scott. "Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship," Cambridge, Massachusetts: The MIT Press, 2005.

Tomko, S., Harris, K. T., Toth, A., Sanders. J., Rudnicky, A, & Rosenfeld, R. "Towards Efficient Human Machine Speech Communication: The Speech Graffiti Project," ACM Transactions on Speech and Language Processing, Vol. 2, No. 1., February 2005.

"Voice eXtensible Markup Language (VoiceXML™) version 1.0," W3C, May 05, 2000. <http://www.w3.org/TR/voicexml/>

Bolchini, D., Colazzo, S., Paolini, P., Vitali, D. "Designing Aural Information Architectures," Proceedings of the 24th annual conference on Design of communication SIGDOC '06. ACM Press, October 2006.

Walker A. M., Di Fabbrizio, J., Mestel, C., Hindle, Don. "What Can I Say ?: Evaluating a Spoken Language Interface to Email," Proceedings of the SIGCHI conference on Human factors in computing systems CHI '98, ACM Press/Addison-Wesley Publishing Co., January 1998.

Perugini, S., Anderson J., T., Moroney, F. W. "A Study of Out-of-turn Interaction in Menu-based, IVR, Voicemail Systems," Proceedings of the SIGCHI conference on Human factors in computing systems CHI '07, ACM Press, April 2007.

Brandaghan, J. Russell, ed. "Design by people for people: Essays on Usability," Usability Professionals Association, 2001, p. 109.

Norman, A., Donald. "The Design of Everyday Things," New York: Basic Books (Perseus), 2002.