Full Writeup

Introduction

Voice user interfaces (VUIs) rely on a particularly social aspect of human communication: speech. Research shows that people react in the same way to synthesized voices as they do to human voices; therefore, "voice interfaces are intrinsically social interfaces" (Nass and Brave 4). Reeves and Nass also propose that since humans are experts at social interaction, they would also become experts at computer interfaces that are designed according to social principles (8). This research looks at how users talk to voice systems and at whether we should model these interactions on conversation.

Background

Voice user interfaces refer to interactive media that use speech as the main or only mode of input and feedback (Harris 3). Such systems have been deployed in a number of domains, including stock trading, airline travel, and banking (Cohen, Giangola and Balogh 9). For mobile users, voice portals allow access to multiple services such as weather information, stock quotes, and directory assistance.

Speech applications offer several benefits for users. They are a natural way of interacting, drawing on the user's own language skills. They can also be accessed anywhere there is a telephone and are especially useful in situations where the user's hands or eyes may be busy, such as driving. For small mobile devices, speech offers an ideal solution for inputting and retrieving complex information (Cohen et al. 11).

In general, speech interfaces can be divided into three categories based on how much control the system has over the dialogue (Zue and Glass 1166). These are user-initiative, system-initiative, and mixed-initiative. In user-initiative systems, users can say whatever they want, although this means that they may be uncertain as to what their options actually are. At the other end are system-initiative systems, which constrain the user's response with questions such as, "Please say just the departure city" (Zue and Glass 1166). Mixed-initiative, or conversational systems lie between the two. These are goal-oriented dialogues in which users and the computer interact as in a conversation.

Dialog in voice interfaces

Harris characterizes human-to-computer dialogue as "an emergent genre" whose characteristics are only recently being developed: "We don't yet know how people prefer to talk to machines" (Harris 21). Still, Harris believes that while people won't speak

to computers exactly as they speak to humans, they will prefer to draw on their natural linguistic and conversational skills to do so (21).

There is already research in the social sciences which shows that people respond to computer-generated voices the same way that they do when interacting with other people (Nass and Brave). The voice of a speech system triggers stereotypes, influences behavior, and affects people's attitudes toward products. People also experience the effects of similarity attraction and consistency attraction to computer voices. In fact, the human brain has evolved to become skilled at recognizing social cues, and this process is automatic (Nass and Brave 3).

However, there are differences between human-human dialog and human-computer dialog. Human-computer talk is likely to be more abrupt, with fewer social pleasantries than in human-human conversations, and the dialogue acts might include "other sequences of identification and role verification," depending on the domain (Harris 102). People could also conceivably be more tolerant of errors while dealing with voice systems (Harris 22).

The balance of participation is also different. A study in an air-travel domain found that in the human-computer exchanges, the system dominated the conversation, talking more and taking the initiative more often (Doran, Aberdeen, Damianos, and Hirschman). In the human-human dialogue, the balance was more equal. The reason for the imbalance could be by design. Doran et al. speculate that "poor speech recognition performance" meant designers were trying to keep users from straying outside the system's capabilities (155).

Meanwhile, Weegels found that one reason users did not speak more was that they simply had a limited understanding of the system's capabilities to handle longer queries (78). Weegels' study also found that while talking to the system, users fell back on their habits and prior experience with human conversation, using other services in the domain, and using computers (75). Patterns of behavior they carried over from human-human interaction included politeness, over-articulation, asking questions, and interjecting (Weegels 79). We can see from this study that users, particularly new ones, draw from their knowledge of analogous interactions when dealing with voice systems, and one of these is human-human dialogue (habits, after all, are hard to break).

But while studying human-computer dialog is important, Harris argues that it has limited value in informing "optimal design principles," as such interaction is necessarily constrained by current technology - so it is likely to change as technology changes - and by speakers' willingness to adapt to unnatural language patterns (22). The model we are left with, then, is conversation.

For system designers, there are issues in modeling human-computer dialog on conversations between humans. According to Zue and Glass, many speech applications are "exercises in information access and/or interactive problem solving," with the both sides working incrementally toward a solution (1166). Human dialog is not the most efficient model for such problem-solving purposes, as it contains fragments, disfluencies, overlaps, and interruptions (Table 1). Some researchers therefore advocate a more structured approach for better task success. But arguably, "users may feel more comfortable with an interface that possesses some characteristics of a human agent" (Zue and Glass 1166, 1167). Also, even small utterances such as *okay* serve a purpose, in this case acknowledgment, so they could also be included in VUI dialogs (Zue and Glass 1167).

Table 1. Transcript of a conversation between an agent A and a client C (Zue and Glass 1167)

C: Yeah, [umm] I'm looking for the Buford Cinema.	Disfluency
A: OK, and you want to know what's showing there or …	Interruption
C: Yes, please.	Confirmation
A: Are you looking for a particular movie?	
C: [umm] What's showing?	Clarification
A: OK, one moment …	Back-channel
A: They're showing a Troll in Central Park.	
C: No.	Inference
A: Frankenstein.	Ellipsis

Regardless of these differences, designers can draw on the principles of conversation to craft a more natural and intuitive interaction. The rules and expectations inherent in human conversation are largely unconscious but violations lead to "interfaces that feel less comfortable, flow less easily, are more difficult to comprehend, and are prone to more errors. Effective leverage of shared expectations can lead to richer communication and streamlined interaction" (Cohen et al. 8).

A particular challenge of writing for a voice interface is simply in working in the mode of speech. Speech is much less formal than writing - compare the stilted "You *must* say your pin number … ," to the much more conversational, "Go ahead and say you pin number … " (Cohen et al. 156). Not only that, speech is ephemeral and builds context "as it proceeds"; writing, because it is usually meant to be read later, tends to have most of the context built in (Harris 434). Speech also contains paralinguistic cues such as tone and volume, for which there are no ready equivalents in writing (Harris 434). In

applying the principles of conversation to writing prompts we can address some of these challenges.

At its root, we see that the dialogue is a discourse occurring within a particular context (Cohen et al. 135). According to Grice's Cooperative Principle, conversation is a mutual endeavor based on shared assumptions and goals, so each utterance should be "such as required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which [it occurs]" (Grice, cited in Harris 79). Grice's conversational maxims (quantity, quality, relevance, and manner) can serve both as "output-writing guidelines" for VUI writers and as "inferential guidelines" to predict what users are going to say (Harris 79).

Cohen et al. divide the language of the interface into two dimensions: cognitive and social (169). The cognitive aspect includes using cohesion devices such as discourse markers and pronouns to aid comprehension, placing new or important information at the end of a sentence, and being aware of the direction of pointer words, which is different in spoken and written English. For the social element, the language should send the right cues through register, or level of formality, which can affects how users view the company's brand. Factors such as word choice and the social role of the persona with regard to the user should be consistent throughout (Cohen et al. 168, 169).

In particular, discourse markers (e.g. *first, next, actually, oh*) act as important conversation management tools, helping to impart context and a sense of progression to the exchange (Cohen et al. 143). Consider the wording of the following prompts, which are devoid of discourse markers:

> "Please say the date.
> Please say the start time.
> Please say the duration.
> Please say the subject."
>
> (Giangola, cited in Cohen et al. 140)

Such an exchange sounds completely unnatural. By adding discourse markers and pronouns (shown in italics), users know where they are in the process:

> *"First*, tell me the date.
> *Next*, I'll need the time *it* starts.
> Thanks. <pause> *Now*, how long is *it* supposed to last?
> *Last of all*, I just need a brief description … "
>
> (Giangola, cited in Cohen et al. 141)

Areas of further research

While this paper has focused on the dialog aspect of the voice interface, the other aspect is its underlying technology. Zue and Glass outline many developmental challenges that must be met before conversational interfaces can be widely deployed. If the goal is a more natural system then speech synthesis should be improved, particularly for "the encoding of prosodic and possibly paralinguistic information such as emotion" (Zue and Glass 1176). Zue and Glass note that "the speech synthesis component is the one that leaves the most lasting impression on users – especially when it does not sound especially natural" (1175).

For this relatively new genre, we might also ask, "How *do* people prefer to talk to computers?" While we know how people react to voice interfaces as they do to people, the question remains as to whether they will also prefer to speak to them as they do to people.

References

Cohen, M., Giangola, J., Balogh, J. (2004). *Voice User Interface Design*. Boston: Addison-Wesley.

Doran, C., Aberdeen, J., Damianos, L., Hirschman, L (2003). Comparing Several Aspects of Human-Computer and Human-Human Dialogues. In J. van Kuppevelt & R.W. Smith (Eds.), *Current and New Directions in Discourse and Dialogue* (pp. 133-159). Netherlands: Kluwer Academic Publishers.

Harris, R.A. (2004). *Voice Interaction Design: Crafting the New Conversational Speech Systems*. San Francisco: Morgan Kaufmann

Nass, C. and Brave, S. (2005). *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship*. Cambridge, MA: MIT Press.

Reeves, B. and Nass, C. (1996). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Stanford, CA: CLSI Publications.

Weegels, M.F. (2000). Users' Conceptions of Voice-Operated Information Services. *International Journal of Speech Technology 3*(2), 75-82. Retrieved February 27, 2007 from SpringerLink.

Zue, V. and Glass, J. (2000). Conversational Interfaces: Advances and Challenges. In *Proceedings of the IEEE, 88*(8). 1166-1180.