Can Voice User Interfaces Say "I"?
An Experiment with Recorded Speech and TTS

Amy Huang, Francis Lee, Clifford Nass
Department of Communication,
Stanford University
Stanford, CA 94305
{amyable, square, nass}@stanford.edu

Young Paik and Luke Swartz
Department of Symbolic Systems
Stanford University
Stanford, CA 94305
CA 94305
{lswartz, ypaik}@stanford.edu

ABSTRACT
How do people respond to voice user interfaces (VUIs) that use first-person pronouns as opposed to those that use thepassive voice? We addressed this question in the context of a phone-based auction via a 2 (first person vs. passive voice) X 2 (recorded vs. synthesized speech) between-participants experiment (N=48). There were significant cross-over interactions with respect to user relaxation, perception of system quality, and bidding behavior such that personal pronouns were responded to more positively than passive voice for recorded speech, while passive voice was responded to more positively for synthesized speech. We discuss implications for the theory and design of VUIs.

Keywords

Voice user interfaces, first-person pronoun, TTS (text-to-speech), consistency effect, CASA (Computers Are Social Actors)

INTRODUCTION

Recently, voice synthesis and recognition technologies have become robust enough to be deployed in a wide variety of products and contexts. While these voice user interfaces (VUIs) may be implemented on GUI computers, they may require very different design strategies.

One of the two key design issues for any VUI is the choice of *prompts*, that is, the words that are spoken by the system. (The other key aspect, speech recognition, is beyond the scope of this paper). In essence, designers must make two decisions about prompts: 1) *What* should be said, and 2) *How* should they sound?

There are a number of books that suggests guidelines for the writing of prompts and supporting text [1, 4, 6]. Unfortunately, most of the advice is based on intuition, anecdote, and taste. Because there is not a systematic body of evidence from which to draw, both because of the rapid changes in technology and the paucity of experimental research, designers are confronted with authors that frequently and vehemently disagree with each other, with no solid evidence with which to resolve the dispute. While everyone agrees that VUIs should not just rattle off a list of options at every point in the interaction, there are disputes about virtually all other wording choices: long or short, repetitive or variable, proactive or reactive, chatty or formal, flattering or factual, etc..

First person in interfaces

In this paper, we address one of the most fundamental issues in the writing of prompts: When should a voice user interface speak in the first person active voice, that is, say "I," "me," "my," etc., or should the system exclusively use third person, passive language, such as "here it is"?

In traditional textual or graphical interfaces, the first person is almost universally eschewed, and there is evidence that the first person is off-putting [12]. The objectivity of computers, the argument goes, should be manifested in their language, just as academic and journalistic writing is normally written in the third person. The mixed message of objectivity and subjectivity presents the user with an inconsistency that undermines the credibility and likeability of the system.

One can also make an ethical argument against the use of "I" in computer-based systems. The use of first-person pronouns suggests agency and subjectivity: Computers do not have agency, so it is unethical to have programmers and designers seemingly remove themselves from responsibility for the behavior of their systems [3]. Furthermore, if individuals do confuse computers with people [11], then it is unethical to encourage that behavior in the language that is selected.

On the other side, many argue that infusing interfaces with "personality" is a plus. The argument is that when life-like entities interact with a user, whether real or synthetic, it seems awkward to avoid first person pronouns. Under this view, interactants who use the third person when talking with someone else implies superiority over or even contempt of their interaction partner. The inconsistency, in contrast to the earlier view, occurs when a rich representation of humanness does *not* say "I." Furthermore, this position argues, the active voice is more

psychologically engaging and more readily processed, as the issue of agency is made explicit rather than implicit. That is, infusing an interaction with "personality" is a plus.

To address this question directly, one must create identical interactions imparting identical content, differing only in whether the interface uses the active or passive voice. That is our approach in the present paper.

Recorded Speech versus Synthesized Speech

While there is heated debate concerning issues of *what words* an interface should use, there is no debate about *"who"* should be delivering the words: Recorded speech is universally agreed to be superior to synthesized speech (TTS). TTS suffers from problems of clarity and prosody eliminated with a well-chosen recorded voice. Most germane to the present study is that TTS does not sound *human* or natural; TTS suggests "machine," even when clarity is not a problem.

While we certainly examine main effects for recorded speech vs. TTS, our primary focus is on *how* the type of speech *interacts* with the use of the first-person or passive voice in VUIs. Are responses to first-person prompts the same regardless of modality, or does the "human-ness" of recorded speech and the "machine-like" character of TTS change the nature and meaning of the prompts?

Consistency Theories

The debate concerning whether to use first-person or passive voice has traditionally been framed as an "all-or-nothing" decision. That is, those researchers who believe that anthropomorphism is a mistake argue that "I" should *always* be avoided; conversely, those who believe that making an interface as human as possible urge the use of first-person whenever possible. Thus, both positions view prompt selection as *independent* of the use of recorded speech or TTS.

Recent research has suggested that a main-effects approach to this decision may be misguided. Specifically, some researchers have begun to argue that *consistency* is a fundamental user concern [9]. Under the consistency argument, each modality should exhibit the *same* degree of humanness [8]. Under this view, then, recorded speech should use "I," while obviously non-human TTS should speak in passive voice.

EXPERIMENT

This study was conducted in the context of a telephone auction site, followed by an online questionnaire. We chose an auction context for three reasons. First, an auction context was chosen because of the proliferation and popularity of on-line auction sites (indeed, almost half of our participants had previously participated in an online auction.). Second, the time-sensitivity and terse content of auction sites make them a natural for telephone-based interfaces. Finally, the auction context permits a straightforward behavioral measure (discussed below).

Questions

Based on the above discussion, the experiment was designed to address the following research questions.

Q1: Does the use of first-person pronouns in a voice interface affect users' evaluations of the interface, the voice, their feelings during interacting with the interface, and their behavior?

Q2: Does TTS and recorded speech affect users' evaluations of the interface, the voice, their feelings, and their behavior?

Q3: Does type of speech interact with prompt style, or do they exert independent effects?

Method

Participants

Students in a communication class at a university ($\underline{N} = 64$) were selected to participate in the study. To ensure comprehensibility of the synthesized speech, all participants were native English speakers. Participants were randomly assigned to conditions, with gender balanced across conditions. Each participant was sent a recruitment email containing the URL for the experiment website and a password. All participants were debriefed at the end of the course, and received class credits for their participation.

Procedure

The experiment was a 2 (with or without first-person pronouns) by 2 (TTS vs. recorded speech) balanced, between-subjects design. Participants were asked to log on to the experiment website from a computer with nearby telephone access. They were instructed not to use call-waiting or speakerphone features while participating in the experiment.

Upon registering at the website, participants were given a scenario in which they are about to graduate and move to another city, and must furnish their new apartment. This scenario was chosen because it was potentially relevant to all university students, regardless of their gender and personal interests. Participants were then instructed to call the auction system, where they would place bids on items.

During the auction process, participants listened to descriptions of five auction items futon, refrigerator, microwave oven, television, and phone/answering machine one at a time. After listening to each item description, participants placed bids on them by speaking to the system over the phone. The system recorded their bids for analysis. For every auction item, a retail price was given in order to provide an anchor point for participants' bids, so as to reduce individual differences. The auction items were chosen to have the following characteristics: 1) they were typical items for furnishing a new apartment; and 2) the stated retail prices were roughly $150 (to stabilize bidding across items and between subjects).

After hearing all five item descriptions, participants were presented with a web-based questionnaire, which asked them to evaluate the auction system, the voice used in the system, and to express their feelings during the interaction.

Manipulation

The CSLU Toolkit was used to present the spoken script for recorded and TTS voices. Participants either heard descriptions spoken by a recorded voice or a TTS voice. To control for idiosyncrasies of the voices used, two versions of both voices were randomly assigned to users. For the recorded speech conditions, we recorded two different male voices. Two versions of the TTS condition were created by manipulating the voice parameters of the default male voice in the Toolkit. (Although both voices were unambiguously male, one had a higher pitch, pitch range, and speech rate than the other).

For personal vs. impersonal speech, descriptions were written either using or not using first-person pronouns. To ensure that the sentences were grammatical as well as not awkward, we made additional changes in the syntax. Despite these syntactic changes, the semantics of the sentence—including the amount and type of information given—was the same for all subjects. The following is an example.

Personal: "The first item I have for you today is a cozy twin-size pine frame futon. It's a great piece of furniture for any room, and very convenient when your friends come over..."

Passive: "The first item is a cozy twin-size pine frame futon. It's a great piece of furniture for any room, and very convenient when friends come over…"

Measures

Self-reported dependent measures on participants' evaluation of the system, the voice, and their feelings during the interaction were obtained using a web-based textual questionnaire. Participants were asked: "How well do each of the following adjectives describe your feelings about [the system, yourself]?" Each question was followed by a series of adjectives; each adjective was addressed on an independent, ten-point Likert-type scale. The scales were anchored by "Describes very poorly" and "Describes very well." A number of indices were created based on theory and factor analysis. All indices were reliable.

Humanness of the auction system was measured by a single questionnaire item, which asked the participants to indicate how much the auction system was "like a person."

Enjoyability of the interaction was an index composed of seven items: annoying (reverse-coded), boring (reverse-coded), engaging, enjoyable, fun, interesting, likable (alpha = .92).

System's usefulness was an index composed of four adjectives: easy-to-use, useful, frustrating (reverse-coded), and complicated (reverse-coded) (alpha = .74).

Feeling relaxed was an index composed of four adjectives: relax, calm, uneasy (reverse-coded), and comfortable (alpha = .92).

In addition to attitudinal measures, we also assessed the participants' behavior. Specifically, we calculated each participants' average bid across the five auction items.

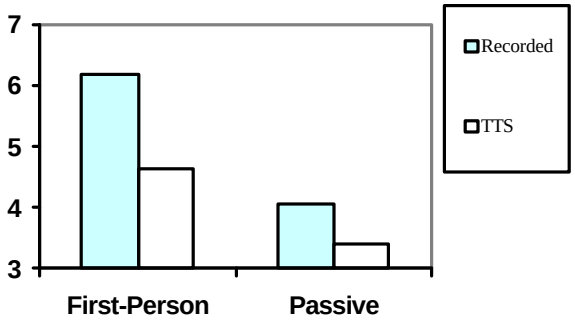
All analyses were based on 2x2 full-factorial ANOVAs. Control of user's experience with on-line auctions did not have a substantive effect on any of our results, so we do not report those analyses here.

RESULTS

Humanness

In order to ensure that we can address consistency issues, we had to ensure that both recorded speech and the first-person were more human than their counterparts. Consistent with this hypothesis, the system using first-person pronouns, when compared with the system using passive voice, was evaluated as having a higher degree of human-ness, $F(1, 63) = 11.38$, $p < .01$. Similarly the system using recorded speech was also evaluated as more like a person, $F(1, 63) = 5.02$, $p < .05$. The interaction was not
significant.

Figure 1: Humanness of the System



Enjoyment of the System

Recorded speech participants enjoyed the system more than TTS participants, $F(1, 63) = 4.38$, $p < 0.05$ (see Figure 2). There was no effect for personal vs. passive and no interaction.

Usefulness of the System

As suggested by the desire for consistency hypothesis, there was a significant cross-over interaction with respect to perceived usefulness of the system, $F(1,63) = 12.29$, $p < .001$ (see Figure 3), such that matching the humanness of the voice and the prompts leads to greater perceived utility. There was no main effect for either modality or personal vs. passive.

Figure 2: Enjoyment of the System

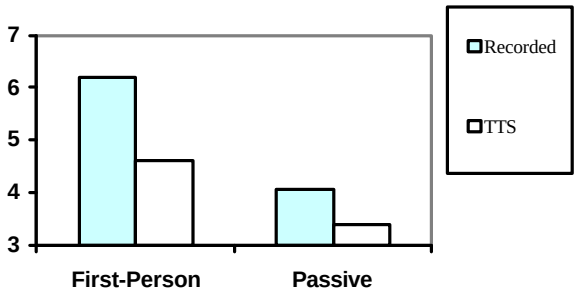
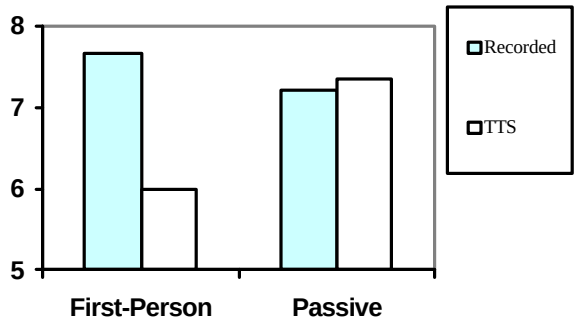


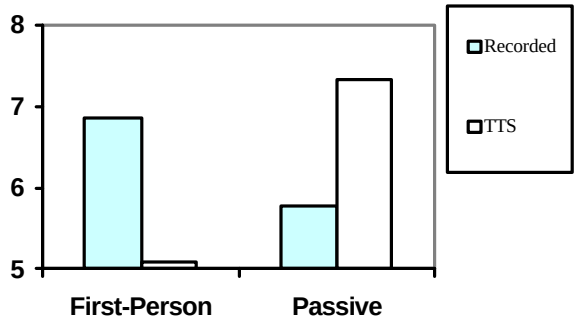
Figure 3: Perceived Usefulness of the System



Feelings of Relaxation

Users are also more relaxed when the humanness of voice and content is matched, $F(1,63)= 19.94$, $p < .001$ (see Figure 4). Specifically, recorded speech users feel more relaxed when encountering first-person language, while TTS participants prefer passive voice. There were no main effects for modality of voice nor for voice content.

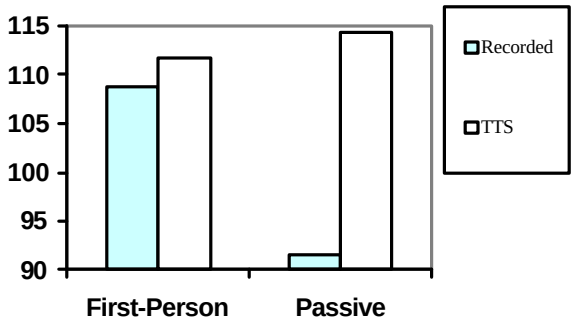
Figure 3: Relaxed Feeling of the User



Bidding Behavior

Finally, there was a significant interaction with respect to the critical behavioral measure, the amount bid, $F(1,63) = 4.04$, $p < .05$ (Figure 5). Again, the interaction pattern fits the consistency hypothesis, as participants bid more when TTS was combined with passive voice, and when recorded speech used personal pronouns.

Figure 5: Amount Bid



DISCUSSION

Traditionally, issues of prompt design and issues of selecting recorded vs. synthesized speech have been seen as independent. Books on prompt design have made absolute statements either for or against the use of "I" and the active voice. Researchers in speech synthesis have argued for using the most human-like-sounding speech possible, regardless of content, and have bemoaned the inadequacies of text-to-speech, including its tendency to sound "unnatural."

The present research suggests that the picture is much more complicated. Rather than users having an absolute preference for "human-like" or "machine-like" interfaces, it is the *consistency* between content and voice that is important. Consistency had implications for perceived quality of the system, how relaxed the user was, and most interestingly, for individual's bidding behavior. That is, designers should pair human-like voices with human-like, personal scripts (after all, most humans use "I" and "me" quite often), while they should combine machine-like voices (i.e., synthesized speech) with less human-like, passive scripts.

One important theoretical implications of this research is that the idea of anthropomorphism may in fact be a dichotomy rather than a continuum. The present research suggests that making an interface "partially human" is not at the midpoint of an underlying dimension; instead, it seems that there are three categories: "human," "not human," and "other," with the latter as clearly undesirable. This is consistent with psychological findings that humans do not like ambiguous categories [2, 5] (this explains the pervasiveness of stereotyping, for example).

On the other hand, the argument by the Computers are Social Actors paradigm [11] that humans automatically respond socially to computers must be modified. Users perceive synthesized speech and passive voice as *less* human than recorded speech and first-person, respectively. Thus, the description of computers as "social actors, just like humans," is too extreme. This is particularly the case when the level of humanness on one dimension conflicts with the level of humanness on another dimension.

Another important design point is that socialness or humanness is not always preferred. In the present case, the interaction effect demonstrates that a system with TTS and passive voice (a highly non-human combination) is no worse than a system with both recorded speech and first-person pronouns (a highly human combination). Thus, one must account for the strong conscious aversion towards humanlike computers as well as the consistency effect, rather than assume that making an interface more human-like will improve its reception.

It is impossible to vary the use of first-person pronouns without also varying some other aspects of speech, such as sentence structure, the use of words, etc. In the present study, two aspects of the scripts changed along with the use or non-use of first-person pronouns: 1) when first-person pronouns were used, the second-person pronouns also appeared more frequently, and 2) when first-person pronouns were avoided, passive voice sentences appeared more frequently. Although these two aspects of the scripts may be considered as independent from the use of first-person pronouns, thus constituting contaminating factors in our study, we consider them as concomitant aspects of the use of first-person pronouns. That is, in the real world, the use of first-person pronouns is associated with certain other aspects of speech. Therefore, eliminating such covariation would be extremely awkward as well as undermine external validity. Of course, further research should examine different ways to instantiate the use of first-person pronouns. Sentence structure and use of second-person pronouns can also be studied independently if they are found to be of theoretical or practical interest.

This study suggests a number of avenues for future research. First, one could test the generalizability of the present findings by exploring other roles for the voice. For example, in an auction, the system is perceived as trying to persuade users and sell them products. In this case, objectivity and credibility may be highly valued, and the user may be concerned with a consistent picture. On the other hand, in entertainment contexts, there may be a more limited depth of processing. In these cases, consistency may be less important, and the main effects of prompt style and modality may hold sway. Indeed, the present results show that for enjoyment, main effects were the only relevant factors.

Another important area is the use of "we" as opposed to "I." Individuals who refer to themselves (with the exception of royalty!) do not use the terms "we" and "us," but individuals representing companies often do use the first-person plural. It is possible that as a marker of a corporation, "we" would be more palatable when spoken by TTS, and perhaps less palatable when spoken in recorded speech (because, like passive voice, it would suggest a diminution of responsibility).

Although this study focused on the perennial question of "I" vs. passive voice, there are many other markers in language that indicate "humanness." Included in this list are emotion [10], metaphor [5], intuition [5], humor [7], etc.. How these aspects of language should be mixed with output type is an open question.

There are also other modalities that suggest humanness, and raise issues of consistency. Should synthetic characters and faces speak in recorded speech, synthetic speech, or text only? [8]. Are semi-anthropomorphic representations, such as the Paperclip in Microsoft Office, disturbing because of inconsistency?

Finally, from a technical standpoint, as the quality of synthesized speech improves, when will it be good enough to be perceived as human? That is, if synthetic speech can mimic the prosody, clarity, and emotion of human speech, while still retaining some indicator that it is machine-generated, the consistency effect may disappear, and it will seem perfectly normal for a computer to say, "The next item I have for you is a great pine-frame futon…." In general, the question is which features of synthesized speech are germane to the judgment "human or not?"

The interaction between voice characteristics and message content also suggests that when comparing the efficacy of various recorded and TTS voices, one must ensure that the content include both first-person and passive voice content; otherwise, the content may bias the perception of the voices, as particular voice characteristics become more or less germane.

As voice user interfaces become more ubiquitous, the number of questions about how to word prompts will grow dramatically. The present research suggests that the answers to these questions will be conditioned by *who* is reading the prompt.

References

1. Balantine, B., Morgan, D. P., Meisel, W. S. 1999. *how to build a speech recognition application*. New York: Enterprise Intergration Group.
2. Fiske, S.T. & Taylor, S.E. 1991. *Social cognition.* New York: McGraw-Hill.
3. Friedman, B. (Ed.). 1999. *Human values and the design of computer technology (CSLI Lecture Notes, No. 72)*. New York: Cambridge University Press/CSLI.
4. Gardner-Bonneau, D. 1999. *Human factors and voice interactive systems*. Dordrecht, Netherlands: Kluwer Academic.
5. Lakoff, G. & Johnson, M. 1983. *Metaphors we live by.* Chicago: University of Chicago Press.
6. Markowitz, J.A. 1995. *Using speech recognition: A guide for application developers.* Englewood, NJ: Prentice-Hall.
7. Morkes, J., Kernal, H. & Nass, C. (2000). Effects of humor in task-oriented human-computer interaction and computer-mediated communication: A direct test of SRCT theory. *Human-Computer Interaction, 14*(4), 395-435.

8. Nass, C. & Gong, L. 1999. Maximized modality or constrained consistency? *Proceedings of the AVSP 99 Conference*, Santa Cruz, CA.

9. Nass, C. & Lee, K. (submitted). Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied*.

10. Picard, R. 2000. *Affective computing*. Cambridge, MA: MIT Press.

11. Reeves, B. & Nass, C. 1996. *The media equation: How people treat computers, television, and new media like real people and places*. New York: Cambridge University Press/CSLI.

12. Shneiderman, B. 1997. *Designing the user interface: Strategies for effective human-computer interaction*. Menlo Park, CA: Addison-Wesley.