Shall We Mix Synthetic Speech and Human Speech? Impact on Users' Performance, Perception, and Attitude

Li Gong
Department of Communication
Stanford University
Stanford, CA 94305-2050 USA
+1 650 814 4690
ligong@stanford.edu

Jennifer Lai
IBM Research
30 Saw Mill River Road
Hawthorne, NY 10532 USA
+1 914 784 6515
lai@watson.ibm.com

ABSTRACT

Because it is impractical to record human voice for ever-changing dynamic content such as email messages and news, many commercial speech applications use human speech for fixed prompts and synthetic speech (TTS) for the dynamic content. However, this mixing approach may not be optimal from a consistency perspective. A 2-condition between-group experiment (N = 24) was conducted to compare two versions of a virtual-assistant interface (mixing human voice and TTS vs. TTS-only). Users interacted with the virtual assistant to manage some email and calendar tasks. Their task performance, self-perception of task performance, and attitudinal responses were measured. Users interacting with the TTS-only interface performed the task significantly better, while users interacting with the mixed-voices interface *thought* they did better and had more positive attitudinal responses. Explanations and design implications are suggested.

Keywords

Mixing Human Speech and Synthetic Speech, Consistency, Speech Applications, Telephone-based Solution, Virtual Assistant, Email and Calendar

INTRODUCTION

Speech technology holds tremendous promise for pervasive computing and is becoming widely incorporated in a range of interfaces and products. The recent years have witnessed a boom in telephone-based speech applications, which range from delivering one's email messages to browsing the World Wide Web via voice. Enabling users to have access to business data such as email messages and calendar entries by dialing a server from any telephone is an important application area. An example of such an application is a speech-based virtual-assistant system that users interact with on the telephone to manage email messages, voicemail messages, and calendar entries. It may be also capable of retrieving information such as stock quotes and news, searching for a restaurant by area code, or assisting with phone calls. There already exist commercial products featuring virtual assistants, for example, Portico by General Magic [13] and the Personal Virtual Assistant by Conita Technologies [12].

An immediate issue in the design of such a speech interface is what type of speech should be used. Two types of speech are available today: recorded human speech and computer-synthesized speech. The latter is also known as text-to-speech (TTS). There are also two types of content in the output in virtual assistant-like interfaces: fixed prompts and dynamic text such as email headers and email bodies. Because it is impractical and too time-consuming, if not impossible, to record human speech for all the dynamic content, there are two approaches to designing speech output in virtual assistant-like interfaces. One approach is to mix human speech and TTS, while the other approach is to use TTS only.

Most commercial applications on the market today adopt the mixing approach. A recorded human voice, often of a voice talent, states the fixed prompts while TTS is used for the dynamic text. An example of such a mix could be: "You have an email message from *Paul Lance* with subject '*Kid sick, out tomorrow morning"*, where the text in italics would be spoken with TTS. The body of the email message would be read exclusively by TTS. Thus, the mixing of human speech and TTS is often within a sentence as well as between sentences.

The assumption behind the mixing approach is to use natural human speech whenever possible. Most likely this is because users report that TTS sounds unnatural and is unpleasant to listen to [10]. TTS lacks both clarity and prosody of normal human speech [9]. The recent developments in concatenative speech synthesis [e.g., 14, 15] improve the naturalness of the sound, but it is still not as good as natural human speech, especially a professional voice talent. Thus mixing TTS and natural human speech appears to be an "optimal" solution. This solution is rooted in a more fundamental approach where the basic premise is to maximize technological excellence for every component or aspect of an interface. This approach is often assumed and followed by technologists and practitioners. Technological optimization is traditionally an ultimate goal in the technology field. When the best technology or choice cannot be applied to every aspect of an interface, a designer may intuitively use the best technology for whichever aspect it can be applied to. In essence, the technological maximization approach is to use the best choice one has for every aspect of an interface.

In contrast to the technological maximization approach, a consistency approach posits that it is important to keep different aspects of an interface consistent with each other. Because of the drastically different quality and nature of human speech and TTS, the frequent mixing of these two types of speech is very likely to cause inconsistency in the interface. It may result in a disjointed interface rather than a consistent and coherent one. Users may find back-and-forth mixing of the two types of speech jarring and hard to adjust to. In consequence, the inconsistency caused by such a mixing may hinder the users' processing and perception of the interface and their interaction with it. Therefore, mixing human speech and TTS is not considered the optimal solution from the consistency approach.

For human-human interactions, consistency has long been evidenced as a general and important rule in the psychology literature [2]. For example, people prefer consistency in another person's personality and consistency among another's communication channels such as verbal and nonverbal channels [1, 3]. Following Reeves and Nass's (1996) Media Equation theory that posits that the users' interaction with computers and interfaces follows social rules [11], the consistency rule would equally apply to interfaces and human-computer interaction. Two recent studies support the consistency theory in the cases of matching the face and the voice in a talking head in an interface [8] and matching personality cues in posture with personality cues in verbal content of a stick figure in an interface [5].

In line with the consistency approach, an alternative solution to mixing TTS and human speech is using TTS for both fixed and dynamic content. Although TTS is more difficult to understand than human speech, a TTS-only interface may maintain consistency in the interface and facilitate the user's smooth interaction with the interface.

To empirically test these two competing approaches and their corresponding predictions for designing virtual assistant-like speech interfaces, an experiment was conducted. The interfaces were evaluated according to users' task performance, self-perception of task performance, and attitude towards the interface. This study not only provides important guidelines for the design of virtual assistant-like speech interfaces but also sheds light on understanding how users respond to and interact with interfaces in general.

METHOD

Experimental Design

A two-condition between-group experiment was conducted where the type of speech used for the output in a virtual-assistant application varied between the groups. In Condition 1, the virtual-assistant system spoke with male TTS throughout. In Condition 2, the virtual-assistant system had two voices. The first voice was the recording of a male voice talent and was used for all the fixed prompts. The second voice was the male TTS voice, which read the dynamic content (e.g., the email header and body). The male TTS used in both conditions was produced by the same TTS engine with identical speech parameters. Although many existing commercial systems [e.g., 12, 13] mix the voice of a female voice talent with male TTS, such a combination mixes the gender of the voice in addition to the type of speech. As the current study focuses on mixing the type of speech used for the voice output, the effect of mixing gender will be examined in a follow-up study.

Participants

Participants were 24 employees (12 males and 12 females) at the IBM T. J. Watson Research Center in New York. To avoid any potential difficulty in understanding the synthetic speech, all participants were native English-speakers with no reported hearing problems. Participants received $20-worth gift certificate or lunch vouchers for their participation. Each of the participants was randomly assigned to one of the two conditions.

Based on participants' responses to a post-experiment questionnaire, there were 9 participants (37.5% of the sample) in the 21-35 age range, 11 (45.8%) in the 36-50 range, and 4 (16.7%) at the age of 51 or above. In terms of education, 4 people (16.7%) reported having Bachelor's degree, 9 people (37.5%) reported master's degree, and 11 people (45.8%) reported doctoral degree. Regarding prior exposure to TTS, 18 participants (75%) reported to have heard TTS once or twice, while 6 people (25%) reported to listen to it with some regularity but less than a few times a week. None of the participants reported working with TTS.

Procedure

The participants took the study one at a time in a usability lab. The participants were told that the purpose of the study was to test a prototype virtual-assistant system. Prior to the start of the study, consent for videotaping was obtained from each participant. Upon arrival in the lab, the participants were seated and given a booklet with the instructions on the first page. The instruction page described the purpose of the study and explained that to date only the email and calendar functions had been implemented in the system. The participants were instructed that their task was to interact with the virtual assistant on the telephone to manage several email and calendar tasks.

They were told that they were not allowed to take notes because the study was intended to maximize a hands-free telephone situation.

The participants were given a background scenario in which "*Paul Lance*" was their manager, and "*Bob Elliott*" was the technical person on their team. In the scenario, it was 8:00 o'clock in the morning on Tuesday August 22, 2000. (The study was conducted in July, 2000.) They were away from their office and dialing the system to check their email messages and calendar. The experimenter assured the participants that all the data collected would be confidential. After the experimenter left the room, the participants dialed the number of the system. They used the telephone on the table in front of them and used the speaker phone so that the system's speech output could be captured in the videotapes.

The booklet also had page-by-page instructions which guided the participants for each involved task. There were eight specific tasks organized around six email messages. One email required updating the calendar in addition to creating an email reply. Another email required only updating the calendar. A third email involved sending an attachment in the email reply. Each page in the booklet contained a brief instruction for the specific task. The instructions were written in such a way that they told the participants what to do, but not how to do it. The following is a sample instruction:

"After you listen to the urgent email, check/modify your calendar accordingly. Please be sure to get back to the person who sent you the email."

The instructions were not so detailed that all participants would perform the tasks with the same steps and verbal input. Command words such as "reply" or "forward" were avoided in order to capture what the participants would say naturally. At the same time, the instructions were sufficient enough so that the participants would know what they were supposed to achieve.

After each task, there was a set of questions asking the participants to evaluate the task they had just completed. Upon completing a task, the participants asked the virtual assistant to "take a break" to put the system on hold and then answered the task-evaluation questions in the booklet. Once they finished answering the questions, they would say "come back" to reactivate the system and proceed to the next email/task. The experimenter gave an overview of this process to the participants before the start of the experiment. On the bottom of each page, there was also a brief instruction reminding the participants what to do next.

After completing all the tasks with the virtual assistant, the participants completed a post-experiment questionnaire. After the questionnaire, the experimenter entered the usability lab to debrief them. A study session lasted around 40 minutes.

System

The experiment used a Wizard of Oz system instead of a working system with speech recognition. The reason was to avoid uncontrollable speech recognition errors. Since this study focused on the impact of mixing types of speech in speech output of the system, we did not want it to be confounded by speech recognition difficulties.

To make the Wizard of Oz system sound realistic to the participants, repair prompts were played in response to complex or unclear input spoken by the participants. Two incremental repair prompts were available for the Wizard to use:

"Sorry, I didn't understand you. Can you say it again?"
"Sorry, I still didn't understand you. You may try rephrasing your request. Thanks."

These prompts were played when judged appropriate by the Wizard rather than at the same point of the interactive task for each participant. The reason is that a repair prompt played in response to an unclear or complex user request suggests a higher quality about the system than the same prompt played in response to a simple and easily recognizable request. The experimenter played the role of the Wizard in the control room adjacent to the usability lab. The participant could be seen through a one-way mirror and heard through an audio system connecting the usability lab and the control room. The Wizard played the correct prompt based on the input spoken by the participant. None of the participants suspected that they were not interacting with a real speech system.

Manipulation

The TTS engine used in the study was IBM Via Voice Outloud. The average speed was 175 words per minute. For all other speech parameters, the default setting was used. IBM Via Voice Outloud was chosen because of its convenience and availability in the location where the study was conducted. It is very unlikely that use of this particular TTS engine would cause any idiosyncrasy in the study because Lai, Wood, and Considine (2000) found no significant difference in comprehension of synthetic speech among this engine and four other commercial engines [7].

A professional male voice talent was hired to record the human voice prompts. Because of the effect of splicing the recorded human prompts with the TTS, the duration of email headings and calendar listings was longer in the mixed-voices condition than in the TTS-only condition. But the length of the body of the email message was identical because it was read by TTS in both conditions. Table 1 lists the mean duration of email messages (including the header) and calendar listings (for the whole given day) in the two conditions.

	Email messages	Calendar listings
TTS-only	22.66	17
Mixed-voices	25.00	31

Table 1: The mean duration of email messages and calendar listings (in seconds)

Measures

Two types of measures were used in the study: the behavioral measure of participants' task performance as rated by coders, and participants' perception and attitude as measured in paper-and-pencil questionnaires.

Behavioral Measure: Task Performance

For task performance, two coders independently reviewed the videotapes of participants. After an overview of the videotapes of all participants, a 0-3 rating scale was constructed to capture the range of the performance of the participants in the study. Based on this scale, the coders independently rated each participant on eight tasks (five email tasks, two calendar tasks, and one attachment task). The denotations of the scale were slightly different for email tasks and calendar and attachment tasks.

For all the tasks, "0" indicates that the participants "did nothing", and "1" indicates "did something, but something wrong". For email tasks, "2" indicates "completed the task but with difficulty *or* with minimum involvement", and "3" indicates "completed the task with ease *and* with high involvement". For calendar and attachment tasks, "2" indicates "completed the task with substantial difficulty" and "3" indicates "completed the task quite easily".

Compared to calendar tasks, email tasks were relatively easy to complete. Only some participants had some difficulty in completing the email tasks. For example, they needed multiple attempts to get the task done. Instead, the participants showed noticeable differences in their involvement level in completing the email tasks. Therefore, involvement level was captured in the rating of email tasks. "Minimum involvement" indicates the participant invested very little thought and/or action in completing the task. For example, an email response of "ok, I got your message" would be "minimum involvement". By contrast, a response with "high involvement" would be clearer or more detailed. Thus, if a participant completed the task with either difficulty or minimum involvement, he or she would receive a rating of "2". Receiving a "3" would require both ease and high involvement in completing the email tasks. Involvement was not relevant to calendar and attachment tasks. Thus only the difficulty level was captured in rating the calendar and attachment tasks.

The inter-rater reliability between the two coders was a high .88. An index of overall task performance was created by averaging the ratings of all the eight tasks, Cronbach alpha = .71.

In addition, the number of times the participants repeated the email messages and calendar listings was also recorded.

Questionnaire Measures: Perception and Attitude

Two sets of paper-and-pencil questionnaire data were collected. After each task, the participants answered the following questions:

- "How well do you think you performed the task?"
- "How well do you think the virtual assistant performed?"

They also rated "completing the task with the virtual assistant on the phone" on four pairs of semantic differential adjectives:

- difficult-easy,
- uncomfortable-comfortable,
- inconvenient-convenient,
- inefficient-efficient.

All these questions were answered on 1-10 scales and all strongly loaded on one single factor in factor analysis for all the tasks. An index of self-perception of task performance was composed of these six questions and created by averaging across all the tasks. Cronbach alpha for this index was .93.

The post-experiment questionnaire consisted of attitudinal questions regarding the virtual-assistant system, the voice(s) of the virtual assistant, and the user experience. Participants' demographic information was collected at the end of the questionnaire. All the questions except the demographic ones were measured by asking how well certain adjectives described the system, the voice(s), and the experience on 1-10 scales ("0" = "describes very poorly", "10" = "describes very well"). An index of ease of use was created regarding the virtual-assistant system and consisted of two items: easy to use and difficult (reverse coded), Cronbach alpha = .85.

Because in the mixed-voices condition there were two drastically different voices, it would be confusing and imprecise to ask questions about "the voice" of the virtual assistant. Thus, the same set of voice-evaluation questions was asked separately for the human voice and the TTS in the mixed-voices condition. The human voice was worded as the "voice in the system that steered the interaction". The TTS was worded as the "voice reading the email messages". This differentiation was not necessary in the TTS-only condition because there was only one voice. Thus, the TTS was just worded as the "voice of the virtual assistant". Two indexes about the voices were constructed through factor analysis:

1) *Clarity of the voice*: consisted of "articulate", "clear", "hard to understand" (reverse coded), "incomprehensible" (reverse coded), α = .92;
2) *liking of the voice*: consisted of "annoying" (reverse coded), "enjoyable", "friendly", "frustrating" (reverse coded), "likeable", "pleasant", and "warm", α = .88.

For the user experience, an index of effort was created and consisted of "challenged", "effortless" (reverse coded), "exhausted", and "strained", α = .75.

RESULTS

T-tests were run to test whether there were differences on the dependent measures between the two conditions. For the behavioral measure of task performance, participants in the TTS-only condition performed the tasks overall significantly better (M = 2.22) than those in the mixed-voices condition (M = 1.68), t(22) = 3.20, $\underline{p}$ < .01. For individual email and calendar tasks, the differences between the performances of the two groups were in the same direction, i.e., the TTS-only group had higher performance than the mixed-voices group. Hence, individual tasks were not differentiated in the analysis and only overall task performance was used in the further analysis. Figure 1 presents the mean difference in overall task performance.

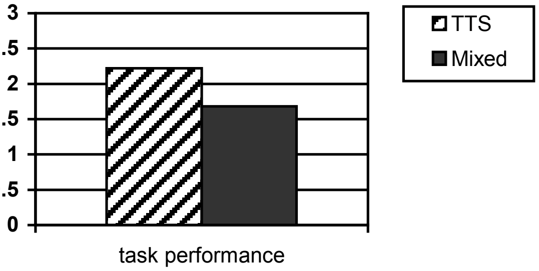

Figure 1: Comparison of means in overall task performance

In terms of repetition of email messages and calendar listings, the participants in the TTS-only condition had significantly more repetition per message or listing (M = 1.01) than those in the mixed-voices condition (M = .41), t(22) = 3.87, $\underline{p}$ < .001. Figure 2 shows the mean difference in repetition. Repetition was not significantly correlated with the overall task performance.

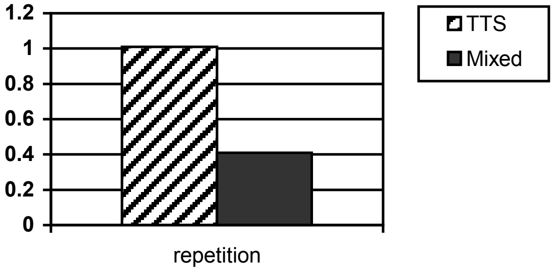

Figure 2: Comparison of means in the average repetition of email messages and calendar listings

In contrast to the between-group difference in coder-rated task performance, participants' self-perception of their task performance was significantly lower in the TTS-only condition (M = 5.38) than in the mixed-voiced condition (M = 6.89), t(22) = 2.24, $\underline{p}$ < .05. Participants in the mixed-voices condition also thought the virtual-assistant system was easier to use (M = 6.58) than those in the TTS-only condition (M = 4.41), t(22) = 2.39, $\underline{p}$ < .05. Figure 3 presents the mean differences in self-perception of task performance and ease of use of the system.

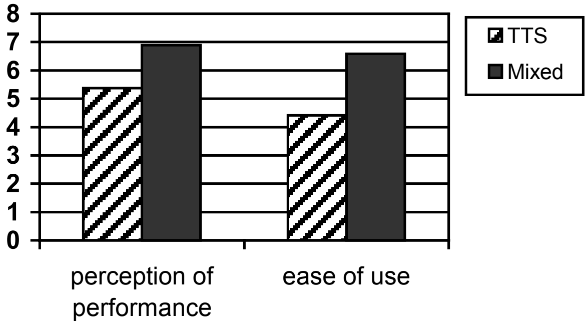

Figure 3: Comparison of means in self-perception of task performance and ease of use of the system

For clarity and liking of the voice, two pairs of comparisons were made through t-tests. First, t-tests compared the TTS in the TTS-only condition and the human voice in the mixed-voices condition. Then, t-tests compared the TTS in the TTS-only condition and TTS in the mixed-voices condition. Although TTS was produced by the same TTS engine with identical parameters in both conditions, they were treated as two voices because they were in two different conditions and had slightly different roles. The TTS in the TTS-only condition read all the texts, while the TTS only read the texts of dynamic content in the mixed-voices condition. The role of TTS in the TTS-only condition also differed from the role of the human voice in the mixed-voices condition because the human voice only read the fixed prompts. The differences in the functions of the voices may hurt their comparability. Thus, caution is taken in interpreting the results about the voices.

Participants thought the human voice was clearer (M = 7.82) than the TTS in the TTS-only condition (M = 2.64), t(22) = 8.18, p < .001; and liked the human voice more (M = 5.01) than the TTS in the TTS-only condition (M = 3.14), t(22) = 4.38, p < .001. Figure 4 presents the mean comparison in clarity and liking of TTS in the TTS-only condition and the human voice. Clarity was significantly and positively correlated with self-perception of task performance (r = .59, p < .01) and ease of use of the system (r = .62, p < .01). Liking was also significantly and positively correlated with self-perception of task performance (r = .58, p < .01) and ease of use of the system (r = .64, p < .01).

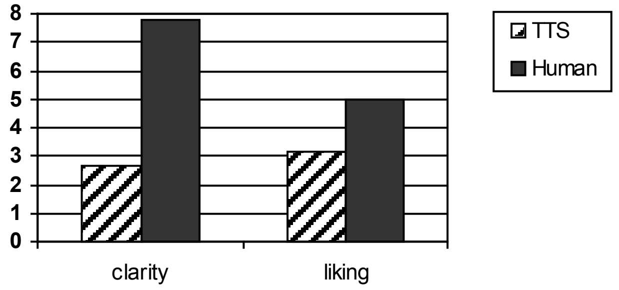

Figure 4: Comparison of means in clarity and liking of TTS in the TTS-only condition and the human voice

In the comparison between TTS in the TTS-only condition and TTS in the mixed-voices condition, participants thought the TTS in the TTS-only condition was clearer (M = 2.64) than the TTS in the mixed-voices condition (M = 1.61), t(22) = 2.63, p < .05; and liked the TTS in the TTS-only condition more (M = 3.14) than the TTS in the mixed-voices condition (M = 1.32), t(22) = 5.73, p < .001. To reiterate, the TTS in the two conditions was produced by the same TTS engine with identical speech parameters and the message content in the two conditions was identical. Figure 5 presents the mean comparison in clarity and liking of TTS in the TTS-only condition and the TTS in the mixed-voices condition.

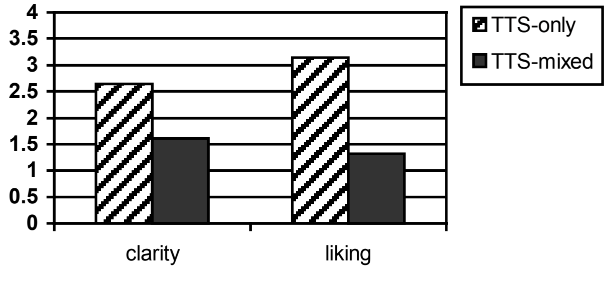

Figure 5: Comparison of means in clarity and liking of TTS in the TTS-only condition and the TTS in the mixed-voices condition

In terms of user experience, participants in the TTS-only condition reported to have put more effort in doing the tasks (M = 5.80) than those in the mixed-voices condition (M = 4.68) at an approaching-significance level, t(22) = 1.82, p = .08. Figure 6 presents the mean comparison in effort.

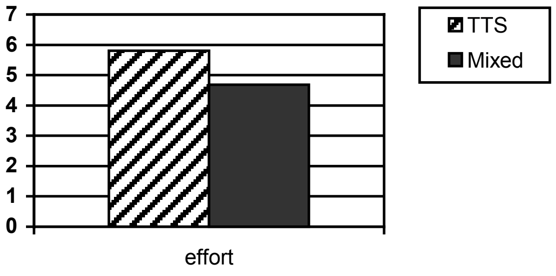

Figure 6: Comparison of means in effort

With respect to demographic variables, no significant differences were found for gender, age, education, or prior exposure to TTS on the task performance, perception or attitudinal measures. Chi-square tests showed no significant difference in distribution of age, education and prior-exposure to TTS in the two conditions. Thus the two groups appeared to be homogenous demographically.

DISCUSSION

As the results have shown, the mixing of TTS and human voice had opposite effects on task performance vs. self-perception of task performance and attitudinal responses. Users had poorer task performance when they interacted with the mixed-voices virtual-assistant interface than when they did with the TTS-only interface. Although users interacting with the TTS-only interface also had a greater number of repetitions of the email messages and calendar listings, the repetition was not significantly related to task performance. Thus, the explanation that people performed better because they listened to the messages more frequently is ruled out. The longer average duration for email messages and calendar listings in the mixed-voices condition is very unlikely to cause the difference in the task performance either because one would expect that a longer duration would give users more time to process and thus lead to better task performance. Instead, consistency of the interface appears to be a reasonable explanation. The TTS-only interface seems to be more consistent and able to facilitate the users' interaction with the interface than the mixed-voices interface.

With the TTS-only interface, users only need to deal with one type of speech and thus may be more able to stay focused on the task. This is supported by our observations of the videotapes. Most of the participants in the TTS-only condition looked very focused and absorbed most of the time during their interaction with the virtual assistant. On the contrary, we observed that participants in the mixed-voices condition who had been sitting back in the chair when listening to the human voice would often lean abruptly forward towards the telephone when TTS started to play. This seems to suggest that the switch in processing the two drastically different voices is costly for the user. Processing one type of speech consistently, even though TTS is relatively difficult to understand, may also enhance a training effect in that one gets better at understanding TTS when he/she hears it more. Unfortunately because of the design of this study, the training effect could not be empirically differentiated from the overall consistency effect.

Moreover, with the TTS-only interface, users also seemed more likely to persist in getting the task done. This is supported by the findings that the participants in the TTS-only condition were more willing to repeat the email messages and calendar listings and reported to have put more effort in doing the tasks than those in the mixed-voices condition.

Interestingly, although TTS was produced by the same TTS engine with identical speech parameters in the two conditions, it was perceived more negatively when it was mixed with the human voice. The contrast with the almost impeccable voice of the voice talent probably made the TTS sound worse. Since TTS “reads” the dynamic content which is crucial for completing the tasks, the more negative perception of TTS in the mixed-voices interface may contribute to the users’ worse task performance.

Provocatively, users who interacted with the mixed-voices virtual-assistant interface *thought* they performed the task better and thought the virtual-assistant system was easier to use than users who interacted with the TTS-only interface. Tentative evidence suggests that this may have been caused by the strong presence of the pleasant voice of the voice talent that steered the interaction with the user. Such a pleasing human voice probably makes the users feel better overall. This voice-preference explanation is partially supported by the positive relationships between liking and perceived clarity of the voice and self-perception of task performance and perceived ease of using the system. A causal relationship cannot be claimed here because of the limitation of the study. During the debriefing after the study, some participants in the mixed-voices condition commented that it was the messages not the short fixed prompts that were hard to understand and needed to be read by a clearer voice. Some participants in the TTS-only condition also commented that the fixed prompts were quite easy to understand. This important user insight seems to further suggest that the human voice may mainly make users feel more pleasant rather than help them better understand the content of the messages and carry out the tasks.

Although the explanations proposed here are suggestive rather than conclusive, the findings of the study demonstrate the importance of examining interfaces from other perspectives in addition to technological maximization. Consistency is an important concept that needs to be extensively examined. Consistency also seems to be a subtle concept that may not be readily measured by explicitly asking users what they would prefer because it was the coder-rated behavioral measure of task performance that supported the consistency prediction in this study. This study also suggests that the quality and pleasantness of the voice in speech interface is a major factor affecting users’ perception and attitude. Inspired by the findings of this study, two design implications are suggested below.

Design Implications

Consistency of the interface is an important principle for interface design. When consistency conflicts with technological maximization, trade-off has to be made. Consistency should be given strong consideration, especially if a consistent interface makes users perform better and the technological maximization does not target the aspect which needs technological improvement most, for example, dynamic content in the virtual assistant case.

Certainly, users’ perception and attitude are also important. Therefore, when designing speech interfaces, the quality and pleasantness of the voice should also be considered important. Within the consistency framework, one should make the voice as good as it can get. The concatenative TTS seems to hold promise for improving the naturalness and pleasantness as well as comprehensibility of TTS. But research is needed to empirically assess whether this promise holds.

Future Research

If concatenative speech meets the promise of providing a more natural, pleasant, and comprehensible TTS, it will be worth testing whether concatenative TTS would increase users’ perception and attitude as well as task performance in their interaction with virtual assistant-like interfaces. If it does, the tension between consistency and technological maximization would be much ameliorated.

To remedy the limitations in the design of this study, future research is needed to enable differentiation of pure training effect of TTS and consistency effect. For example, users could receive substantial training for the TTS prior to the start of the experiment and then be equally assigned to the TTS-only and the mixed-voices conditions. Future research is also needed to further test the suggested explanation that a pleasant and impeccable voice of a voice talent causes the more positive subjective perception in users. One could use a less pleasing but still fluent and clear non-professional human voice to mix with TTS to test this explanation.

As mentioned earlier in the paper, the combination of female human voice and male TTS will be tested in future research because most commercial applications use a female human voice in their systems. The effect of mixing gender in addition to mixing types of speech needs to be examined. As female TTS is improving, it should be included in future research as well.

In the mixed-voices virtual-assistant interface, the reason for why two different types of voices are used is not explained to the users. Although the reason for mixing is obvious to designers and speech technologists, it may not be so to most users. In the social setting of multiple speakers, the Master of Ceremony or the first speaker normally introduces the second speaker. Using this analogy, the human voice in the virtual-assistant application would be the Master of Ceremony because he/she greets the user and steers the interaction, and thus he/she should introduce the second voice, i.e., the TTS. The human voice could, for example, introduce and frame the TTS as a kind of robotic assistant and explain the technological reason for the use of TTS. This reasonable framing of TTS may ameliorate the problem of inconsistency, win users' understanding and positive attitude, and help them prepare for the mixing of the voices. This posited effect of framing the TTS needs to be empirically tested.

Finally, the consistency issue in mixing voices is an important issue for speech interfaces in general. How multiple voices and types of speech should be used in speech interfaces is an important design question as well as an important research topic. Moreover, the consistency concept deserves examination in a range of other interface issues, such as speech input and output, face and voice pairing, emotional expression of an agent and the context, and the design of an interface in relation to the culture it is designed for.

ACKNOWLEDGMENTS

We greatly appreciate Cliff Nass for his precious advice and insight on the study, Marisa Viveros and David Wood for their generous support, Steve Heitmann for his valuable support in technology, Karen Cheng for her very helpful comments on the paper, and anonymous reviewers for their very valuable comments.

REFERENCES

1. Domangue, B.B. Decoding effects of cognitive complexity, tolerance of ambiguity, and verbal-nonverbal inconsistency. *Journal of Personality, 46,* 519-535, 1978.
2. Fiske, S.T. and Taylor, S.E. *Social Cognition*. New York: McGraw-Hill, 1991.
3. Graves, J.R. and Robinson, J.D. Proxemic behavior as a function of inconsistent verbal and nonverbal messages. *Journal of Counseling Psychology, 23*, 333-338, 1976.
4. Hendrick, C. Effects of salience of stimulus inconsistency on impression formation. *Journal of Personality & Social Psychology, 22,* 219-222, 1972.
5. Isbister, K. and Nass, C. Consistency of personality in interactive characters: Verbal cues, non-verbal cues, and user characteristics. *International Journal of Human-Computer Studies, 53,* 251-267, 2000.
6. Kelley, H. H. Attribution theory in social psychology. In D. Levine (Ed.), *Nebraska Symposium on Motivation* (Vol. 15, pp. 192-240). Lincoln: University of Nebraska Press, 1967.
7. Lai, J., Wood, D., and Considine, M. The effect of task conditions on the comprehensibility of synthetic speech. *The Proceedings of CHI '00* (The Hague, The Netherlands, March 2000), ACM Press, 321-328.
8. Nass, C. and Gong, L. Maximized modality or constrained consistency? *The Proceedings of AVSP'99* (Santa Cruz CA, August 1999), 1-5.
9. Olive, J.P. "The talking computer": Text-to-speech synthesis. In Stork D.G. (Ed.), *HAL's Legacy: 2001's Computer as Dream and Reality* (pp. 101-131*)*. Cambridge, MA: MIT Press, 1997.
10. Ralston, J.V., Pisoni, D.B., and Mullennix, J.W. Perception and comprehension of speech. In Syrdal, A.K., Bennett, R.W., Greenspan, S.L. (Eds.), *Applied Speech Technology* (pp. 233-288). Boca Raton: CRC Press, 1995.
11. Reeves, B. and Nass, C. *The Media Equation: How People Treat Computers, Television, and New Media like Real People and Places*. New York: Cambridge University Press/CSLI, 1996.
12. www.conita.com
13. www.genmagic.com/portico/portico_home.shtml
14. www.lhsl.com/realspeak/demo.cfm
15. www.research.att.com/projects/tts