INTERNATIONAL JOURNAL OF SPEECH TECHNOLOGY 6, 123–131, 2003

To Mix or Not to Mix Synthetic Speech and Human Speech? Contrasting Impact on Judge-Rated Task Performance versus Self-Rated Performance and Attitudinal Responses*

LI GONG
SAP Labs, Inc., 3475 Deer Creek Road, Palo Alto, CA 94034, USA

JENNIFER LAI
IBM Research, 30 Saw Mill River Road, Hawthorne, NY 10598, USA

Abstract. Since it is impractical to prerecord human speech for dynamic content such as email messages and news, many commercial speech applications use recorded human speech for fixed content (e.g. system prompts) and synthetic speech for dynamic content. However, mixing human speech and synthetic speech may not be optimal from a consistency perspective. A two-condition between-participants experiment ($N = 24$) was conducted to compare two versions of a telephony application for Personal Information Management (PIM). In the first condition, all the system output was delivered with synthetic speech. In the second condition, users heard a mix of human speech and synthetic speech. Users managed several email and calendar tasks. Users' task performance was rated by two independent judges. Their self-ratings of task performance and attitudinal responses were also measured by means of questionnaires. Users interacting with the interface that used only synthetic speech performed the task significantly better, while users interacting with the mixed-speech interface *thought* they did better and had more positive attitudinal responses. A consistency framework drawn from human psychological processing is offered to explain the difference in task performance. Cognitive processing and attitudinal response are differentiated. Design implications and directions for future research are suggested.

Keywords: mixing human speech and synthetic speech, consistency, text-to-speech, speech interfaces

The two primary types of speech available for delivering spoken output in a speech application are recorded human speech and computer-synthesized speech. The latter, also referred to as text-to-speech (TTS), has improved substantially in recent years but still lags behind natural speech in clarity and prosody (Olive, 1997; van Santen et al., 2000). Many users report that TTS sounds unnatural and is unpleasant to listen to (Ralston et al., 1995). The advantage of TTS is that it can dynamically convert any written text to spoken output. By comparison, recording human speech is both time-consuming and costly. As a result, TTS is often viewed as a more practical option for delivering dynamic content.

The output of a speech application can be categorized into predictable (fixed) content or unpredictable (dynamic) content. A voice talent can be used to record fixed content, such as system prompts, in advance. A range of content such as digits or city names can also be pre-recorded and spliced together as needed to create the prompts. Indeed, most speech applications delivering predictable or fixed content use pre-recorded human speech. Applications that deliver dynamic content, which is not known in advance (e.g. news stories or email messages), tend to use TTS to render the text at the run time.

*An earlier and less detailed version of the paper was presented at the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI' 01), Seattle, WA.

Many speech applications inherently include both dynamic and fixed content. An example is Personal Information Management (PIM) telephony applications that enable users to access their email messages and calendars from any telephone. Recent years have witnessed the growth of this type of application due to the needs of an increasingly mobile work force. An example mixing dynamic and fixed content in this type of application would be: "You have an email message from *Paul Lance* with subject '*Kid sick, out tomorrow morning*"', where the text in italics is dynamic content. Then, an immediate design question is which type of speech to use to deliver content containing both fixed and dynamic texts. A common approach in the speech application industry is to mix human speech with TTS. An alternative approach is to use TTS exclusively.

The mixing approach uses recorded human speech to deliver the fixed text and uses TTS for the dynamic text. In the email example mentioned above, the text in italics would be spoken with TTS while the rest would be spoken with recorded human speech. The body of the email message would be read exclusively by TTS. Thus, the switch between human speech and TTS might be within a sentence as well as between sentences. The premise of this approach is to use the better option, i.e., the human speech wherever possible and use TTS for the rest. In doing so the overall speech quality of the application is posited to be optimal. This point of view is rooted in the traditional philosophy of *technological maximization*, which advocates using the "best of the breed" for applicable components or aspects of an application. Using a better option for a portion of the application is assumed to achieve better quality than using the less good option for the entire application.

In contrast to the technological maximization approach, a *consistency* approach advocates using one type of speech throughout (e.g., only TTS or only human speech). This approach stresses the importance of keeping different aspects of an interface consistent with each other. Because of the drastically different qualities and nature of human speech and TTS, frequent mixing of these two types of speech is likely to cause inconsistency in the interface. For the user, switching between processing these different types of speech may be too cognitively demanding. The interspersed human speech may disrupt the users' process of getting familiarized with TTS, which would be more effective if it is continuous. Although the existence of a pleasant human voice may boost the users' impression of the system, the inconsistency and interference caused by it may adversely affect the user's speech and cognitive processing and as a result their task performance.

In the psychological literature, consistency is widely evidenced as a principle in human's cognitive and social processing. Numerous studies show that simultaneous presentation of inconsistent objects such as a red color patch with the word "blue" creates interference in people's perception and hinders their task performance (Dyer, 1973; Kahneman and Chajczyk, 1983; Stroop, 1935). The perceptual interference caused by inconsistency is often referred to as "Stroop-effect" because of the seminal study conducted by Stroop (1935). Auditory presentation of words with inconsistent vocal attributes such as the word "high" spoken with a low pitch (110 Hz) also created perceptual interference (Hamers and Lambert, 1972). Consistency is also proven essential in perceiving and deriving a person's personality (Kelley, 1967) and in processing a person's verbal and nonverbal cues and comprehending the meaning of the communication (Roy and Sawyers, 1990). The human propensity for consistency can be explained by the Gestalt theory of perception which claims that people inherently need to resolve inconsistency and achieve a coherent perception of a target (Asch, 1946).

A few existing studies in human-computer interaction (HCI) provided evidence in support of the consistency perspective. Consistency in personality cues was found important between the postural cues and the verbal cues of an interface stick figure (Isbister and Nass, 2000) and between the verbal cues and the vocal cues of an interface voice agent (Nass and Lee, 2001). In the visual speech domain, Gong et al. (2001) showed that a computer-synthesized talking head speaking with recorded human speech was less effective in eliciting users' personal information than the talking head speaking with TTS because both the talking head and the TTS are synthetic and consistent with each other. Furthermore, Gong (2001) showed the same consistency effect for a videotaped human face talking with his own speech compared to talking with TTS.

Hence, there seem to be contradicting perspectives with respect to whether it is advantageous to mix human speech and TTS. Two studies examined the effects of mixing human speech and TTS on users. Spiegel (1997) varied the mix of recorded human speech (male and female) and TTS (male only) in delivering listings in a telephone directory to telephone operators. The greeting text such as "welcome to. . ." was always delivered with recorded human speech (Spiegel, 1997, p. 59). The

carrier phrase such as "the address is. . ." was delivered with recorded human speech in some conditions and with TTS in the other conditions. The listing of the dynamic content such as "123 main street" was delivered with TTS in all conditions. This study showed that operators' transcriptions of listings were more accurate when the carrier phrases were delivered with TTS versus recorded male or female human speech. However, the subjective questionnaire data showed that operators perceived the system output being clearer and preferred the system more when it used recorded human speech for the carrier phrases and TTS for the listings than when the system used TTS for both carrier phrases and listings. The mix of recorded female natural speech and male TTS was the least preferred. A drawback in this study is that since all conditions presented the initial welcome phrase in recorded human speech, there was not a single condition that used TTS exclusively.

McInnes et al. (1999) compared mixing human speech with TTS to using TTS exclusively. They only assessed users' perceived difficulty and overall judgment of the system and did not have any performance measure. Users in the study reported a more positive attitude towards the system that mixed human speech with TTS than the one using TTS exclusively. Thus, in terms of the subjective perception and attitudinal responses, this study appears to resonate with Spiegel's (1997) findings. But this study failed to assess users' task performance.

In an attempt to more extensively examine the issue of mixing human speech and TTS, an experiment was conducted to compare mixing human speech and TTS to using TTS exclusively. In addition to assessing the users' attitudinal responses and subjective perception, the users' performance in completing the tasks was measured through ratings of independent judges.

Method

A two-condition between-participants experiment was conducted. The type of speech used for the output of a telephone-based PIM virtual-assistant system varied between the conditions. In Condition 1, the virtual-assistant system spoke with male TTS exclusively. In Condition 2, the virtual-assistant system had two voices. The first voice was the recording of a male vocal talent that was used for all the fixed texts. The second voice was the male TTS voice, which read the dynamic texts (e.g., the email subject and body). The male TTS used in both conditions was produced by the same TTS engine with identical parameters. Although most commercial systems[1] mix the speech of a female vocal talent with male TTS, such a combination mixes the gender of the voice in addition to the type of speech. Spiegel (1997) showed preliminary evidence of the adverse effects of mixing genders of human speech and TTS. Since the focus of this study was on mixing the type of speech, future research needs to more systematically investigate the effects of mixing gender.

Participants

Participants were 24 employees (12 males and 12 females) in a large computer science research organization in the State of New York of the United States of America. To avoid any potential difficulty in understanding English, all participants were native English-speakers with no reported hearing problems. Participants received gift certificates or lunch vouchers for their participation. The participants were randomly assigned to the conditions. Demographic information about the participants was collected in a post-experiment questionnaire and is summarized in Table 1.

Procedure

The participants took the study one at a time in a usability lab. Prior to starting the study, consent for videotaping was obtained from each participant. The participants were given a booklet with the general instruction and the study scenario on the first page. The experimenter also explained the purpose and the procedure of the study to them. The purpose of the study was framed as testing a prototype virtual-assistant system. The participants were told that only the email and calendar functions had been implemented in the system and their task was to interact with the virtual assistant on the telephone to manage several email and calendar tasks. They were not allowed to take notes because the study was intended to maximize a hands-free speech situation. They were supposedly away from the office and dialing the virtual-assistant system to check their

Table 1. Demographic information about the participants in the study.

Age	21–35: 37.5%	36–50: 45.8%	Over 51: 16.7%
Education	Bachelors: 16.7%	Masters: 37.5%	PhD: 45.8%
TTS exposure	Once or twice: 75%	Some regularity: 25%	Work with it: 0%

email messages and calendar items. The experimenter assured the participants that all the data collected would be confidential. After the experimenter left the room, the participants dialed the number for the system. They used the telephone on the table in front of them and used the speaker mode so that the system's speech output could be captured by the video camera.

The booklet had page-by-page instructions for guiding the participants through each specific 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 but sufficient instruction for the specific task. 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 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 do.

After each task, there was a set of questions asking the participants to evaluate their own performance on that task. The participants used the commands "take a break" and "come back" to pause and reactivate the system, respectively. After completing all the tasks the participants answered a post-experiment attitudinal questionnaire. After the questionnaire, the experimenter entered the usability lab to debrief them. An experimental session lasted approximately 40 minutes.

System

The experiment used a Wizard of Oz method instead of using a speech system with speech recognition. The reason was to avoid uncontrollable speech recognition errors. Since this study focused on the impact of mixing the type of speech in the system's output, we did not want any confounding due to speech recognition difficulties.

To make the interaction appear realistic to the participants, repair prompts were played in response to complex or unclear input uttered 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*."

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 speech prompt in response to the input of the participant. None of the participants suspected that they were interacting with a human wizard.

Manipulation

The TTS engine used was IBM Via Voice (VV) Outloud. The rate of speech was 175 words per minute on average. The default setting was used for all other speech parameters. VV Outloud was chosen for its convenience and availability. It is very unlikely that use of this particular TTS engine would cause any idiosyncrasy in the study because Lai et al. (2000) found no significant difference in comprehension of synthetic speech when comparing this engine and four other major commercial engines.

A professional male vocal talent was hired to record the fixed texts. Due to splicing the recorded human speech with the TTS, the duration of email headings and calendar listings was longer in the mixed-speech 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. Calendar listings, in particular, had extensive splicing which resulted in much longer duration in the mixed-speech condition, due to the frequent intermixture of fixed and dynamic content. Table 2 lists the average duration of email messages (including the header and the body) and calendar listings (for the entire day) in the two conditions.

Table 2. The average duration of email messages and calendar listings (in seconds).

	Email	Calendar listings
TTS-only	22.7	17.0
Mixed-speech	25.0	31.0

Table 3. Description of the scale denotations for the different types of tasks.

	0	1	2	3
Email tasks	Did nothing	Did something, but something wrong	Completed the task but with difficulty *or* with minimum involvement	Completed the task with ease *and* with high involvement
Calendar and attachment tasks	Did nothing	Did something, but something wrong	Completed the task with substantial difficulty	Completed the task quite easily

Measures

Judge-Rated Task Performance. Two judges independently reviewed the videotapes of all participants. A 0–3 rating scale was constructed to capture the range of the performance in the study. The denotations of the scale were slightly different for the five email tasks and for the two calendar and one attachment tasks. See Table 3 for a description of the scale denotations for the different types of tasks.

Compared to the calendar and attachment tasks, the email tasks were relatively easy to complete. Only a few participants had difficulty in completing the email tasks. They occasionally needed multiple attempts to get the task done. Instead, the participants showed noticeable differences in their involvement levels in completing the email tasks. "Minimum involvement" indicates the participant invested very little thought and/or effort in completing the task. For example, an email response of "ok, I got your message" was considered to be minimum involvement. By contrast, a response with "high involvement" would include more details and be more elaborate. 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" in an email task would require both ease and high involvement. Involvement, however, was not relevant to the calendar and attachment tasks. For these tasks, only the difficulty level was captured in the rating.

The judges independently rated each participant for all the tasks. The inter-rater reliability (the correlation between the ratings of the two judges) was a high .88. An index of overall task performance was created by averaging the ratings for all the tasks, Cronbach $\alpha = .71$. In addition, the number of times that the participants repeated the email messages and calendar listings was also recorded.

Self-Rated Performance and Attitude. Self-rated task performance was measured by a set of six items on a paper-and-pencil questionnaire after each task. The first two items consisted of the questions: "How well do you think you performed the task?" and "How well do you think the virtual assistant performed?" The other four items were to rate "completing the task with the virtual assistant on the phone" on four pairs of semantic-differential scales: "difficult-easy", "uncomfortable-comfortable", "inconvenient-convenient", and "inefficient-efficient". All six items were answered on 1–10 scales and strongly loaded on one factor in factor analysis. An index of self-rated task performance was composed of these six items and created by averaging across all the tasks, $\alpha = .93$.

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 ("1" = "describes very poorly", "10" = "describes very well"). An index of the ease of using the system consisted of two items: "easy to use" and "difficult" (reverse coded), $\alpha = .85$.

Because in the mixed-speech condition there were two drastically different voices, it would have been confusing and imprecise to ask questions about "the voice" of the virtual assistant. Therefore, the same set of voice-evaluation questions was asked separately for the human voice and the TTS voice in the mixed-speech condition. The human voice was referred to as the "voice in the system that steered the interaction". The TTS voice was referred to as the "voice reading the email messages". This differentiation was not necessary in the TTS-only condition because there was only one voice, which was just referred to 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), and "incomprehensible" (reverse coded), $\alpha = .92$;

2. *Liking of the voice*: consisted of "annoying" (reverse coded), "enjoyable", "friendly", "frustrating" (reverse coded), "likeable", "pleasant", and "warm", $\alpha = .88$.

For the user experience, an index of effort was composed of "challenged", "effortless" (reverse coded), "exhausted", and "strained", $\alpha = .75$.

Results

T-tests were run to test whether there were differences on the dependent measures between the two conditions. For the judge-rated task performance, participants in the TTS-only condition performed the tasks overall significantly better ($M = 2.22$) than those in the mixed-speech condition ($M = 1.68$), $t(22) = 3.20$, $p < .01$. 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 better performance than the mixed-speech group. Therefore, 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 for 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-speech condition ($M = .41$), $t(22) = 3.87$, $p < .001$. Figure 2 shows the mean difference for repetition. Repetition was not significantly correlated with the overall task performance.

In contrast to the difference in the judge-rated task performance, participants' self-rated task performance was significantly lower in the TTS-only condition ($M = 5.38$) than in the mixed-speech condition ($M = 6.89$), $t(22) = 2.24$, $p < .05$. Participants in the mixed-speech condition also thought the virtual-assistant system was easier to use ($M = 6.58$) than those in the TTS-only condition ($M = 4.41$), $t(22) = 2.39$, $p < .05$. Figure 3 presents the mean differences for self-rated task performance and the perceived ease of using the system.

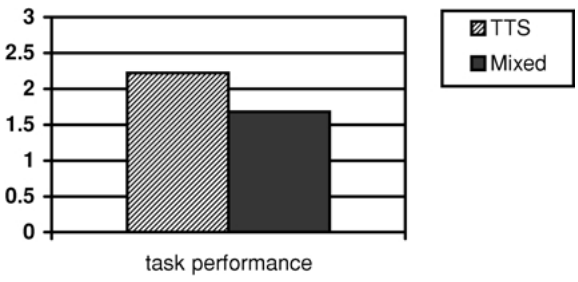

Figure 1. Comparison of means for overall task performance.

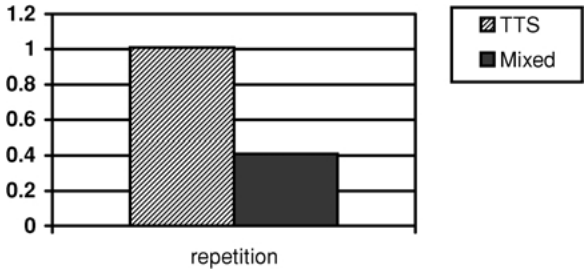

Figure 2. Comparison of means for the average repetition of email messages and calendar listings.

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

Participants thought the human voice was clearer ($M = 7.82$) than the TTS voice in the TTS-only

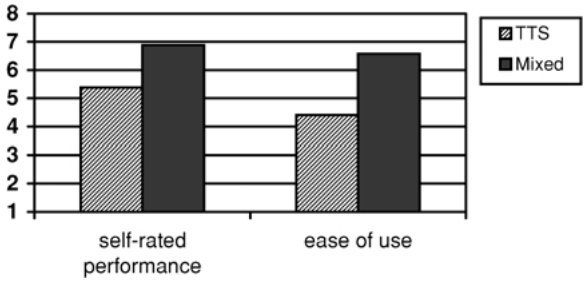

Figure 3. Comparison of means for self-rated task performance and the perceived ease of using the system.

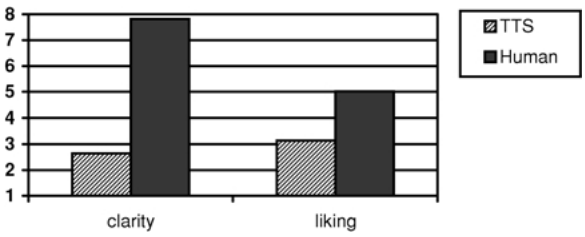

Figure 4. Comparison of means for clarity and liking of TTS in the TTS-only condition and the human voice.

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

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

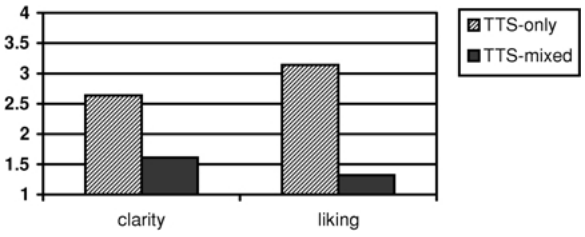

Figure 5. Comparison of means for clarity and liking of the TTS voice in the TTS-only condition and the TTS voice in the mixed-speech condition.

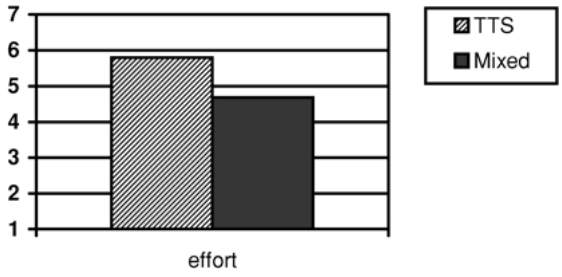

Figure 6. Comparison of means for the perceived effort.

condition ($M = 4.68$) at an approaching-significance level, $t(22) = 1.82$, $p = .08$. Figure 6 presents the mean comparison for the perceived effort. With respect to demographic variables, no significant differences were found on the dependent measures for gender, age, educational level, or prior exposure to TTS.

Discussion

As the results have shown, mixing TTS and human speech had opposite effects on judge-rated task performance vs. self-rated task performance and attitudinal responses. Users had poorer task performance when they interacted with the mixed-speech virtual-assistant interface than when they did with the TTS-only interface. Although users interacting with the TTS-only interface had a greater number of repetitions of the email messages and calendar listings, the repetition was not significantly correlated with task performance. Hence, the explanation that people performed better because they listened to the messages more frequently is ruled out. The longer average duration of email messages and calendar listings in the mixed-speech condition is very unlikely to cause the difference in the task performance because one would expect that a longer duration would give users more time to process the speech and content and lead to better task performance. Therefore, consistency of the interface seems a reasonable explanation. The TTS-only interface is more consistent and conducive to the users' cognitive processing of the speech and consequently their task performance, compared to the interface mixing human speech and TTS.

With the TTS-only interface, users only need to process one type of speech. After they make the initial adjustment to the speech, they are more able to maintain that processing model than if they have to switch between processing the two very different types of speech. A training effect may take place for TTS in that users get more familiar with it and better at comprehending

it when they continuously listen to it. The consistency in speech processing may also help users stay focused and involved in the task. This is supported by the observation of the videotapes that 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 system. On the contrary, the participants in the mixed-speech condition sometimes moved abruptly forward from sitting back in the chair when TTS was played after a human-voice prompt. This strongly suggests there is a switch happening in the processing of the speech and the switch may be disruptive and cognitively costly for the user.

Interestingly, although it was the same TTS in the two conditions, the TTS was perceived more negatively when it was mixed with the human voice. The sharp contrast with the almost impeccable voice of the professional talent probably made the TTS sound worse. Since TTS reads the dynamic content that is crucial for completing the tasks, the more negative perception of TTS in the mixed-speech interface may hurt users' processing of TTS and contribute to their worse task performance.

Provocatively, users who interacted with the mixed-speech interface *thought* they performed the tasks better and thought the virtual-assistant system was easier to use than users who interacted with the TTS-only interface. This resonates with McInnes et al.'s (1999) finding that mixing human speech and TTS elicits more positive attitude from users than using TTS only. Tentative evidence suggests that the more positive self-perception of task performance and attitude may have been caused by the strong presence of the pleasant human voice 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 findings that liking of the voice was significantly and positively correlated with self-rated task performance and the perceived ease of using the system and that the perceived clarity of the voice was also significantly and positively correlated with self-rated task performance and the perceived ease of using the system. A caveat is that a causal relationship cannot be claimed here due to the limitation of the study. During the debriefing after the study, some participants in the mixed-speech condition commented that it was the "real content" (meaning the dynamic content), not the short fixed prompts, that was 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 dynamic content and carry out the tasks.

Theoretical and methodological implications can be derived from this study. First, it is important and fruitful to examine the interface and its effects on users from multiple perspectives. In addition to the traditional technological maximization perspective, consistency deserves equal attention and more research efforts. Second, the classic psychological differentiation between others-observed behavior and self-perceived behavior should receive wide consideration, particularly with respect to the methodology of assessing human-computer interaction. One should be cautious about making general conclusions based on the data collected in only one aspect. This study shows measuring both judge-rated performance and self-rated performance and attitude provides more informative evidence and insight.

The findings of the study also offer implications for interface design. Consistency of the interface is an important design principle. When consistency conflicts with technological maximization, trade-offs have 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 that has the greatest need for technological improvement (e.g., the dynamic content in a spoken output system).

Clearly, users' perception and attitude are also important. When designing speech interfaces, the quality and pleasantness of the voice are highly desirable. Within the consistency framework, one should make the voice as high quality as it can be. The more recent concatenated TTS seems to hold promise for improving the naturalness and comprehensibility of TTS. But research is needed to empirically assess whether this promise holds.

Future Research

If concatenated TTS fulfills the promise of sounding more natural, pleasant, and comprehensible, it will be worth testing whether it would improve users' perception and attitude as well as task performance. If it does, the tension between consistency and technological maximization would be much ameliorated.

As stated earlier, processing one type of speech consistently may foster a training effect in that one gets

better at understanding the speech. Although this effect may be inherent in an interface with one type of speech, a future study could have substantial training with TTS prior to the study to determine whether it helps processing TTS, particularly for the condition that mixes TTS with human speech.

Future research is needed to further test the suggested explanation that a pleasant and professional voice causes more positive subjective experience in users. One could use a less pleasant but still fluent and clear non-professional human voice to mix with TTS to test this explanation.

In the mixed-speech interfaces, the reason for why two different types of speech are used is usually 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 previous speaker normally introduces the next speaker. Using this analogy, the human voice in a mixed-speech interface would be the Master of Ceremony because he/she greets the user and steers the interaction and should introduce the second voice, i.e., the TTS. The human voice could, for example, introduce and frame the TTS as a robotic assistant and explain the technological reason for the use of TTS. This framing of TTS might help prepare users for the mixing of the two types of speech. However, research needs to test this speculation and whether the framing would ameliorate the problem of inconsistency.

Finally, consistency is a general factor concerning speech interfaces. Other issues involving consistency include mixing voices of the same type of speech, matching the visual and auditory components in a multi-modal interface, matching speech input with speech output, matching the emotional cues of speech with the emotional tone of the context, and casting of speech style and voice personality in relation to the traits of the users or the characteristics of the cultural context.

Note

1. Examples are Portico virtual assistant by General Magic and the Personal Virtual Assistant by Conita Technologies.

References

Asch, S.E. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology*, *41*:1230–1240.

Dyer, F.N. (1973). Interference and facilitation for color naming with separate bilateral presentations of the word and color. *Journal of Experimental Psychology*, *99*:314–317.

Gong, L. (2001). Pairing media-captured human versus computer-synthesized humanoid faces and voices for talking heads: A consistency theory for interface agents. Doctoral Dissertation, Stanford University, California.

Gong, L., Nass, C., Simard, C., and Takhteyev, Y. (2001). When non-human is better than semi-human: Consistency in speech interfaces. In M.J. Smith, G. Salvendy, D. Harris, and R. Koubek (Eds.), *Usability Evaluation and Interface Design: Cognitive Engineering, Intelligent Agents, and Virtual Reality.* Mahwah, NJ: Lawrence Erlbaum Associates, pp. 1558–1562.

Hamers, J.F. and Lambert, W.E. (1972). Bilingual interdependencies in auditory perception. *Journal of Verbal Learning and Verbal Behavior*, *11*:303–310.

Isbister, K. and Nass, C. (2000). Consistency of personality in interactive characters: Verbal cues, non-verbal cues, and user characteristics. *International Journal of Human-Computer Studies*, *53*:251–267.

Kahneman, D. and Chajczyk, D. (1983). Tests of the automaticity of reading: Dilution of Stroop effects by color-irrelevant stimuli. *Journal of Experimental Psychology: Human Perception and Performance*, *9*:497–509.

Kelley, H.H. (1967). Attribution theory in social psychology. In D. Levine (Ed.), *Nebraska Symposium on Motivation.* Lincoln, NE: University of Nebraska Press, vol. 15, pp. 192–240.

Lai, J., Wood, D., and Considine, M. (2000). The effect of task conditions on the comprehensibility of synthetic speech. *Proceedings of the Conference on Human Factors in Computing Systems (CHI '00)*, The Hague, The Netherlands: ACM Press, pp. 321–328.

McInnes, F.R., Attwater, D.J., Edgington, M.D., Schmidt, M.S., and Jack, M.A. (1999). User attitudes to concatenated natural speech and text-to-speech synthesis in an automated information service. *Proceedings of Eurospeech '99 (European Conference on Speech Communication and Technology)*. Budapest, Hungary, pp. 831–834.

Nass, C. and Lee, K.M. (2001). Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied*, *7*(3):171–181.

Olive, J.P. (1997). "The talking computer": Text-to-speech synthesis. In D.G. Stork (Ed.), *HAL's Legacy: 2001's Computer as Dream and Reality*. Cambridge, MA: MIT Press, pp. 101–131.

Ralston, J.V., Pisoni, D.B., and Mullennix, J.W. (1995). Perception and comprehension of speech. In A.K. Syrdal, R.W. Bennett, and S.L. Greenspan (Eds.), *Applied Speech Technology*. Boca Raton, FL: CRC Press, pp. 233–288.

Roy, L. and Sawyers, J.K. (1990). Interpreting subtle inconsistency and consistency: A developmental-clinical perspective. *Journal of Genetic Psychology*, *151*:515–521.

Spiegel, M.F. (1997). Advanced database preprocessing and preparation that enable telecommunication services based on speech synthesis. *Speech Communication*, *22*:51–62.

Stroop, J.R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*:643–663.

van Santen, J., Macon, M., Cronk, A., Hosom, P., Kain, A., Pagel, V., and Wouters, J. (2000). When will synthetic speech sound human: Role of rules and data. *Proceedings of International Conference of Spoken Language Processing*. Beijing, China, pp. 878–882.