

Does Adding a Synthetic Face Always Enhance Speech Interfaces?

Li Gong

Department of Communication
Stanford University
Stanford, CA 94305-2050 USA
+1 650 968 6685
ligong@stanford.edu

Clifford Nass

Department of Communication
Stanford University
Stanford, CA 94305-2050 USA
+1 650 723 5499
nass@stanford.edu

ABSTRACT

Does a synthetic face enhance the social appeal of TTS-speech interface and recorded-human-speech interface? A 2x2 between-subjects experiment (synthetic face vs. no face by TTS vs. recorded human speech) was conducted ($N = 48$). Results show that the synthetic face enhanced the social appeal of the TTS-speech interface, but decreased the social appeal of the recorded-human-speech interface. A consistency argument is proposed.

Keywords

Talking Face, Synthetic Face, TTS, Recorded Human Speech, Pairing Face and Speech, Consistency, Social Appeal of the Interface

INTRODUCTION

Talking faces have received substantial attention and efforts in computer laboratories and commercial interfaces [8, 9, 14]. The advantage of talking faces is three-fold. First, speech technology and interfaces are continuing growing. Although the advantage of speech is most evident in the auditory-only modality such as telephone, its less requirement of reading ability, natural and social appeal, and eyes-free and hands-free capability also make speech a desirable and appealing feature on visual interfaces.

Second, faces or face-like icons have long tradition of being used in interfaces [7]. Faces are also becoming widely employed in Web sites [e.g., 15, 16, 17]. The premise is that faces enhance the social appeal of the interface and make the interaction with the interface more fun and lively [14].

Third, combining face and speech, i.e., talking faces or visual speech, incorporates the advantages of both speech and faces. Visual speech has long been evidenced to enhance speech processing and perception [9]. Face-to-face

communication is considered the primary mode of communication between humans [1]. A dynamic talking face is considered more lively and fun than still pictures of faces or speech alone. Thus, talking faces hold tremendous promise for enhancing social appeal of the interfaces. Then the next question is what options the current technologies provide for faces and speech?

In the speech arena, there are recorded human speech and computer-synthesized speech. The latter is also known as text-to-speech (TTS). To date, TTS still lacks both the clarity and prosody of natural human speech [12]. Users have reported TTS sounds unnatural and is unpleasant to listen to [4]. But TTS does not need to be pre-recorded and can be readily and spontaneously produced with any text content with low cost and barely any time delay. Thus, depending on the need and content of the interface, both recorded human speech and TTS are used in various speech interfaces.

In the face arena, although both videotaped human face and computer-synthesized face are available, the pre-videotaped natural human face cannot be synchronized with spontaneous speech content. The new technology of textual-mapping, which takes a videotaped human face and maps it on a computer facial model, is yet to provide natural-looking faces [9]. Hence, the computer-synthesized face is the only viable solution for the dynamic talking face, which can produce spontaneous speech content. Synthetic faces now can perform very well in real-time, especially with respect to lip-synchronization and emotion manifestation [9]. But synthetic faces, like TTS, are obviously non-human and readily distinguishable from natural human faces [9].

Then to create a talking face, there are two choices: a synthetic face talking with TTS or a synthetic face talking with recorded human speech. Then an immediate question is:

Does adding a synthetic face enhance the social appeal of both TTS-speech interface and recorded-human-speech interface?

The answer could be “yes”, “no”, or “depends”, depending on the perspective one takes.

The answer would be “yes” if one believes the synthetic face is always a good thing to have on interfaces. From this perspective, synthetic faces are considered to be lively and fun and facilitate speech perception. Thus, a synthetic face would enhance the social appeal of both TTS-speech interface and recorded-human-speech interface.

The answer would be “no” if one believes the synthetic face is always bad for interfaces because it looks unnatural. Instead of making the interface more social and fun, it might actually disturb the users. From this perspective, a synthetic face would decrease the social appeal of both TTS-speech interface and recorded-human-speech interface.

The answer would be “depends” if one takes the consistency perspective. The consistency perspective posits that different aspects of an interface should be consistent with each other. Because recorded human speech is clearly human and synthetic face is clearly non-human, the combination of them would be inconsistent. By contrast, synthetic face and TTS are consistent with each other because they are both clearly computer-synthesized. Hence, the consistency argument would claim that a synthetic face would enhance the social appeal of TTS-speech interface but decrease the social appeal of recorded-human-speech interface.

Although the consistency perspective is a relatively new concept to the study of computer interfaces and human-computer interaction, it has long been evidenced as a general rule governing social interactions in the psychology literature [3]. People prefer to interact with individuals that behave consistently, even if consistently undesirably, as compared to individuals that behave inconsistently [3, 13]. For example, the inconsistency between one’s nonverbal behavioral cues and his/her verbal content suggests lying [2]. The consistency rule would equally apply to interfaces and human-computer interaction, following the “Computers Are Social Actors” paradigm [11, 13]. This paradigm claims that people automatically and unconsciously treat computers and other new media as real social actors and places. Human-computer interaction is fundamentally social. Social rules governing human-human interaction also apply to human-computer interaction, according to the CASA paradigm. Directly testing the consistency rule in interfaces and human-computer interaction, a study has found that users were disturbed by the inconsistency in personality cues between persona posture and textual content of a stick figure on an interface [5]. Hence, following the consistency perspective and the CASA paradigm, face and speech should be consistent.

To put in statistical terms, the predictions from these three competing perspectives go as the following:

- 1) The argument of synthetic face being always good would predict a main effect that a synthetic face would enhance the social appeal of both TTS-speech interface and recorded-human-speech interface;

- 2) The argument of synthetic face being always bad would predict an opposite main effect that a synthetic face would decrease the social appeal of both TTS-speech interface and recorded-human-speech interface;
- 3) The consistency perspective would predict an interaction effect that a synthetic face would enhance the social appeal of TTS-speech interface but decrease the social appeal of recorded-human-speech interface.

METHOD

Experimental Design

To test these competing perspectives, a 2x2 between-group experiment (synthesized face vs. no face by TTS vs. recorded human speech) was conducted. Participants of the experiment interacted with a computer system that asked questions using one of the four modalities:

- 1) a synthetic talking face with TTS
- 2) a synthetic talking face with recorded human speech
- 3) TTS only without face
- 4) recorded human speech only without face

Participants

Participants were 48 undergraduate students enrolled in a large communication course at a West-coast U.S. university. To avoid potential difficulties in understanding TTS, all participants were native English speakers. The participants received course credit for participating in the study. Each of the participants was randomly assigned to one of the four conditions, with gender balanced across conditions.

Procedure

The participants were told the purpose of the experiment was to test a computer-based interviewing system. They completed the experiment one at a time in a media lab. Upon arrival in the lab, they were asked to read the Informed Consent Form and assured that the information they submitted in the study was totally confidential. After they signed the consent form and read the instruction on the computer screen, they completed the practice round with the assistance of the experimenter. The purpose of the practice round was to demonstrate how to: 1) use the mouse to answers questions on a Likert-type scale; 2) type information into a text box; and 3) use the “Submit” and “Repeat” buttons.

After the practice round, the experimenter left the room. In the first round of the computer-based interview, the computer (via the assigned modality) asked a series of 20 standard questions that assessed impression management. After each question, the participants indicated their answers by using the mouse to click on a response button on a 1-7 scale. The second round consisted of nine open-ended

questions that assessed level of self-disclosure. The participants typed their answers in a text box; when done, they clicked the “Submit” button. For both rounds, there was a “Repeat” button that, when pressed, had the computer repeat the question. After they finished the experiment, the participants left the room and were then thanked and debriefed by the experimenter.

Manipulation

The CSLU Toolkit was used to create the stimuli and to run the experiment. For synthetic speech, we used the Festival TTS engine in the Toolkit. For recorded speech, we recorded the voice of an adult American male; we set the critical TTS parameters (speech rate, fundamental frequency, and frequency range) to be similar to the recorded voice.

In the face conditions, we employed the “Baldi” face provided with the Toolkit. The face was placed on the left side of the screen. The face was 17.8 cm high and 12.5 cm wide. We synchronized “Baldi” with both TTS and the recorded speech using the Toolkit. The interfaces were presented on a 43.2 cm diagonal computer monitor connected to a Gateway Destination 2000 system.

Measures

Social appeal of the interface was captured by two types of measures: participants’ tendency of impression management and their self-disclosure. The idea is the greater the social appeal of the interface, the more pressure the users would feel to manage themselves in a socially positive light [14]. Also the greater the social appeal of the interface, the more willing the users would be to disclose about themselves and the more intimate their disclosure would be because revelation of personal information is a highly social act [10].

To measure impression management, the Impression-Management (IM) subscale in BIDR (Balanced Inventory of Desirable Responding) [6] was used. The original items were first-person statements, for example, “I sometimes tell lies if I have to”. To suit the interview nature of this study, the items were adapted to “Do you” or “Have you” questions, such as “Do you sometimes tell lies if you have to?” The original 1-7 Likert-type scale was retained (1 = “not true”, 7 = “very true”). The responses to the 20 BIDR-IM questions were averaged to form an impression-management index (Cronbach’s $\alpha = .80$). A higher value on the impression-management index indicates a greater tendency for impression management. The computer also recorded the time that subjects took in answering each BIDR question. Time-on-task is a particularly useful measure and reflects users’ attention and involvement in the task.

Self-disclosure was measured by Moon’s [10] nine open-ended self-disclosure questions. A sample question is:

“What has been the biggest disappointment in your life?” Two aspects of self-disclosure were captured. The amount of self-disclosure indicated how willing people were to disclose about themselves and was measured by the average number of words in the participant’s responses to the nine questions. The reliability of this index was very high ($\alpha = .85$).

The second aspect was the depth of self-disclosure which indicated how intimate the disclosure was. It was measured by two independent judges rating the depth of the participants’ disclosure on a five-point Likert-type scale (1 = “low intimacy”, 5 = “high intimacy”). The inter-rater reliability was .74; disagreements were resolved by averaging. The assigned value for a given participant was the average depth of disclosure across the nine items; the reliability of the index was very high ($\alpha = .86$). Because the time participants spent on answering disclosure questions would necessarily correlate with the number of words they typed, the time on this task was not captured.

RESULTS

Full-factorial ANOVA’s were conducted on all the dependent measures. Face and speech showed consistent cross-over interaction effects on all of the measures. The consistency prediction was supported.

For Impression Management, there was a significant cross-over interaction, $F(1, 44) = 4.3, p < .05$ (see Figure 1). Participants who interacted with the synthetic face speaking with TTS exhibited greater impression management than those who only heard the TTS without the face. Conversely, the participants who interacted with the synthetic face speaking with the recorded human speech showed *less* impression management than those who only heard the recorded speech without the synthetic face. Thus, the prediction of the consistency perspective is supported in the measure of impression management.

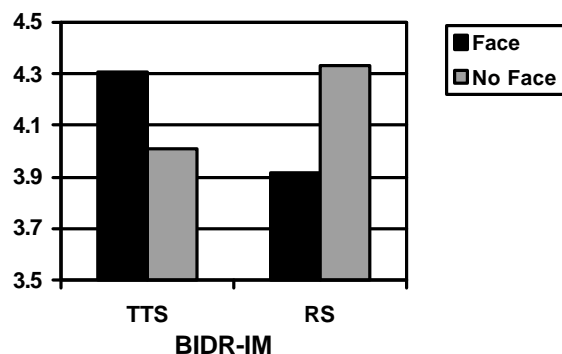


Figure 1: Comparison of means in Impression Management.

The consistency prediction was also observed with respect to amount and depth of self-disclosure. There were

significant cross-over interactions for the amount of self-disclosure, $F(1, 44) = 187.9, p < .001$; and for the depth of self-disclosure, $F(1, 44) = 13.9, p < .001$. Participants disclosed more information about themselves and the disclosure was more intimate when the interface was the synthetic face talking with TTS than when it was TTS alone (see Figures 2 and 3). On the contrary, they disclosed less about themselves and the disclosure was less intimate when the synthetic face was talking with recorded human speech than when the interface incorporated recorded human speech alone (see Figures 2 and 3).

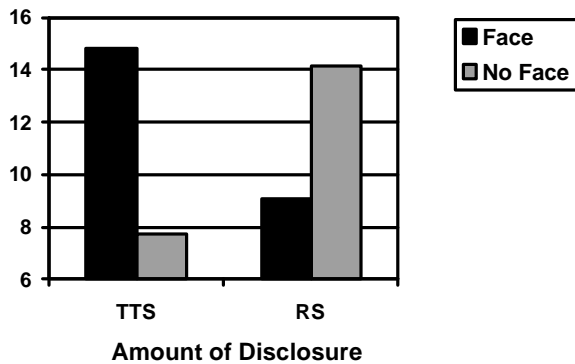


Figure 2: Comparison of means in the amount of self-disclosure.

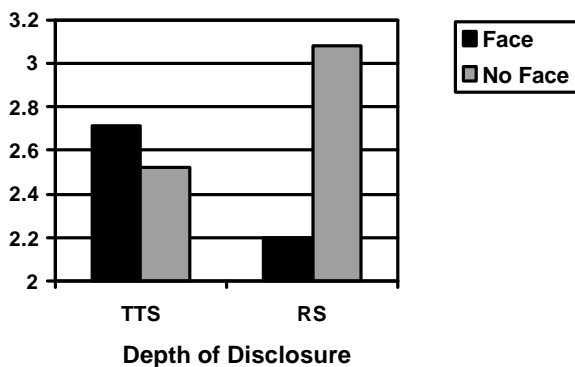


Figure 3: Comparison of means in the depth of self-disclosure.

Further evidence for consistency was found in the significant cross-over interaction for the average time that the participants spent in answering BIDR questions, $F(1, 44) = 12.7, p < .001$. The participants spent more time on BIDR questions when the interface was the synthetic face talking with TTS than when there was TTS alone. Conversely, they spent less time when the interface included recorded human speech with the synthetic face as compared to when the interface was recorded human speech alone (see Figure 4). Also as a main effect, participants spent more time with TTS than with recorded

human speech, a function of greater difficulty in processing TTS, $F(1,44) = 86.0, p < .001$.

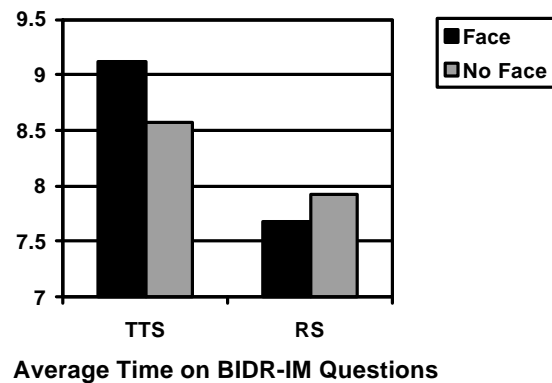


Figure 4. Comparison of average time spent on BIDR-IM questions (in seconds).

To summarize, while a synthesized face consistently enhanced the social appeal of TTS-speech interface, the synthetic face *decreased* the social appeal of recorded-human-speech interface. When the synthetic face was added to the TTS-speech interface, people felt more social with the interface: had greater impression management and more extensive and intimate self-disclosure and spent more time on the task. When the synthetic face was added to the recorded-human-speech interface, people felt less social with the interface: had less impression management and less extensive and intimate self-disclosure and spent less time on the task.

DISCUSSION

The study demonstrates that the consistency between face and speech is important and significantly affects the social appeal of an interface. A synthetic face does not always enhance the social appeal of a speech interface. It does not always decrease the social appeal of a speech interface either. The key is the synthetic face should be consistent with the speech. The synthetic face is consistent with the synthetic speech (TTS), but inconsistent with recorded natural human speech. The reasons are that the synthetic face and the synthetic speech are both computer-synthesized and have obvious machine-like marks. They appear consistent together and enhance each other and the social appeal of the overall interface. On the contrary, the synthetic face and natural human speech do not mix well. Users probably feel confused or disturbed by the combination of clearly-human speech and clearly-non-human face.

This study points to an important implication for interface design. Keeping face and speech consistent is important and consequential. Of course, consistency between face and speech is just one type of pairing in interfaces. More research is needed to assess the consistency issue between

other aspects of the interface. Nonetheless, this study strongly brings up and demonstrates the importance of examining interfaces and human-computer interaction from a consistency perspective.

Because of the nature and scope of the study, limitations exist. To fully test the consistency issue in pairing face and speech, video-recorded human face needs to be included. Although it is not a technologically viable option in designing dynamic talking faces yet, it can be manually manipulated in laboratory settings. The inclusion of natural human face will provide a complete test of the consistency thesis in the pairing of face and speech and provide more guideline for designing future talking-face interfaces.

Also as this study focused on the social aspect of the interface, a study is needed to replicate the design of this study in assessing the speech perception and comprehension of the users. Then we would have a more extensive picture of how the pairing of face and speech affects users' cognition as well as attitude and behavior.

To examine consistency as a general concept in interfaces and human-computer interaction, research is needed to examine the relationships between a range of aspects of the interface. For example, if a computer expresses emotion without being able to detect emotion, will users find that disturbing? Will systems that adapt to the user be expected to have greater intelligence in other domains as well? Must three-dimensional character representations have more fluidity of movement than two-dimensional representations? Should text output be matched with text input and speech output with speech input? As interfaces are becoming more multi-faceted and acquiring more capabilities, consistency between different aspects of an interface will only become a more important issue in research and design.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the enormous technical support provided by David Merrill, Jacques de Villiers, and Jonas Beskow, and the important insights of Dominic Massaro.

REFERENCES

1. Clark, H.H. *Using language*. Cambridge University Press. 1996.
2. Ekman P., and Friesen, W.V. Nonverbal leakage and clues to deception. *Psychiatry*, 32, 88-95. 1969.
3. Fiske, S. T. & Taylor, S. E. *Social Cognition*. New York: McGraw-Hill, 1991.
4. Francis, A.L., Nusbaum, H.C. Evaluating the quality of synthetic speech. In Gardner-Bonneau, D. (Ed.), *Human factors and voice interactive system* (pp. 63-97). Boston, MA: Kluwer Academic Publishers, 1999.
5. Isbister, K. & Nass, C. Personality in conversational characters: Building better digital interaction partners using knowledge about human personality preferences and perceptions. *Proceedings of the WECC Conference, Lake Tahoe, CA*, 1998.
6. Kroner, D. G. & Weekes, J. R. Balanced inventory of desirable responding: Factor structure, reliability, and validity with an offender sample. *Personality and Individual Differences*, 21(3), 323-333, 1996.
7. Laurel, B. Interface agents: Metaphors with character. In Laurel B. (Ed.), *The Art of Human-Computer Interface Design* (pp. 355-365). New York: Addison-Wesley, 1990.
8. Lundeberg, M. & Beskow, J. Developing a 3D-agent for the August dialogue system. *The Proceedings of the International Auditory-Visual Speech Processing Conference* (pp. 151-156), Santa Cruz, California. 1999.
9. Massaro, D. M. *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*. Cambridge, MA: MIT Press, 1998.
10. Moon, Y. Intimate self-disclosure exchanges: Using computers to build reciprocal relationships with consumers. *Working paper for Harvard Business School*, 1998.
11. Nass, C., Moon, Y, Morkes, J., Kim, E-Y, and Fogg, B.J. Computers are social actors: A review of current research. In Friedman B. (Ed.), *Moral and Ethical Issues in Human-Computer Interaction* (pp. 137-162). Stanford, CA: CSLI Press. 1997.
12. Olive, J. P. "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, 1997.
13. Reeves, B. & 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.
14. Sproull, L, Subramani, M., Kiesler, S., Walker, J. H. & Waters, K. When the interface is a face. *Human-Computer Interaction*, 11, 97-124, 1996.
15. www.ananova.com
16. www.mysimon.com
17. www.schwab.com