Human-Computer Interaction -- INTERACT'03
M. Rauterberg et al. (Eds.)
Published by IOS Press, (c) IFIP, 2003, pp. 161-167

Human and Humanoid Don't Match: Consistency Preference and Impact on Users' Trust

Li Gong

SAP Labs, 3475 Deer Creek Rd., Palo Alto, CA, 94304. USA

li.gong@sap.com

Abstract: For many anthropomorphic representations on computer interfaces, there are choices of using media-captured human components and using computer-synthesized humanoid counterparts. This applies to pair the face and the voice of a talking-head character. It is not uncommon that designers mismatch the face and the voice of a talking head in the human vs. humanoid dimension. Based on the literatures on person perception and communication processing, consistency is proposed as a natural human preference which calls for matching the face and the voice. A 2x2x2x2 mixed-design (human face vs. humanoid face by human voice vs. humanoid voice by male vs. female users by attitudinal vs. cognitive task) experiment (N = 80) was conducted. The results supported the principle of consistency preference in terms of users' trust and other attitudes. Attitudinal responses in accordance to consistency preference were stronger among female users than males. For the cognitive task, however, human voice was preferred over the synthetic voice regardless of the face.

Keywords: consistency, trust, human vs. humanoid, face-voice pairing, talking heads, context, gender difference

1 Introduction

For the value of human appeal, liveliness, and intuitiveness, anthropomorphic representations such as talking heads, pictorial characters, and voices have become prevalent on computer interfaces. Talking heads, for example, are used as Web newscasters (www.ananova.com), educators (Massaro, 1998), and receptionists (Lundeberg & Beskow, 1999). For most anthropomorphic representations, there are choices of using media-captured human components such as videos and recorded speech and using computer-synthesized humanoid counterparts such as computer-animated faces and synthetic speech (also called text-to-speech, TTS). An essential question for designing a talking head is then how one should pair the face and the voice in the human vs. humanoid dimension. Although human faces and voices prevail in naturalness and pleasantness, their applications often require manual efforts. Synthetic humanoid faces and voices are still less natural and lower quality but can be programmed to automatically present dynamic content. There exist applications where a human face talks with synthetic speech to deliver dynamic content and a synthetic face talks with recorded speech to achieve speech clarity. In addition to the technical feasibility and quality, does mismatching the face and the voice of a talking head in the human vs. humanoid dimension matter in terms of the effects on users?

A *consistency* perspective argues for the congruency between the face and the voice in the human vs. humanoid dimension. Mixing a human face with a synthetic voice or a synthetic face with a human voice creates inconsistency, violates people's schema of what is human and what is humanoid, and may incur negative effects on users. The consistency perspective is based on the psychological and communication literatures which show that consistency preference is a basic principle in humans' perception of others and social interaction.

1.1 Consistency

Theories and research in social cognition and interpersonal communication state that consistency is preferred in people's perception of others and processing of others' communication cues. Consistency was identified as an important principle in Asch's (1946) Gestalt theory of person perception and Kelley's (1967) personality attribution model. Inconsistency in another person's traits is found to lead to confusion and negative affect and require more cognitive processing. For example, inconsistency perceived in a person's traits was found to lead to disliking of that person (Hendrick, 1972). When a hypothetical person was presented with two inconsistent traits, the participants took longer time to reach judgment on whether the person

would be suitable for an occupation than when the person was presented with consistent traits (Schul, Burnstein, & Martinez, 1983).

Studies on communication processing in social interaction have shown that inconsistency between different communication channels, mainly between verbal and nonverbal channels, adversely affects judgment, attitudes, and interpersonal interaction. For example, inconsistency between a person's verbal and nonverbal cues was found to cause confusion about the person's attitude and intent (Roy & Sawyers, 1990), make the person perceived as less sincere (Argyle, Alkema, & Gilmour, 1971; Graves & Robinson, 1976), and cause greater physical distance from the person (Graves & Robinson, 1976). In particular, inconsistency between verbal and nonverbal cues was demonstrated to be a major indicator of deception (Ekman & Friesen, 1969).

The consistency preference was also found evident in human-computer interaction in three recent studies. One study found that a stick figure in the interface was more persuasive, likeable, useful, and helpful in the Desert Survival Task when its verbal content in a word balloon and its postures were consistent along the extroversion-introversion personality dimension than when they were inconsistent (Isbister & Nass, 2000). Another study found that when a TTS voice agent in a simulated auction site had consistent extroversion or introversion cues in its voice and language, the users liked the descriptions more and found the writers of the descriptions more trustworthy than when the voice agent had inconsistent cues (Nass & Lee, 2001).

As the first attempt to test the consistency issue on the topic of face-voice pairing for talking heads, the author conducted a 2x2 (synthetic face vs. no face by synthetic voice vs. human voice) experiment (Gong, Nass, Simard, & Takhteyev, 2001). The no-face voice-alone conditions were included to control for the main effect of the voice. The effect on users' trust was examined because trust is an important factor in human-computer interaction and prior research has showed that inconsistency contributes to suspicion of deception and hurts the perception of trustworthiness and sincerity. Trust was measured by the amount and the depth of users' self-disclosure in response to open-ended intimate questions asked by the talking head or the voice alone, which was framed as an interviewing agent. Prior research has shown that trust is positively related to self-disclosure (Wheeless & Grotz, 1977). Users disclosed more information about themselves and the disclosure was more intimate when the synthetic face spoke with TTS than with human speech, while human speech without the face elicited more intimate disclosure than TTS without the face. Hence, the pairing of the synthetic face and synthetic speech elicited more trust than the pairing of the synthetic face and human speech, probably because the synthetic face was consistent with the synthetic voice but inconsistent with the human voice.

However, there are limitations in this study that make the conclusion tentative. The main limitation was the lack of a human face in the mix. Two other limitations are the gender of the users and the nature of the context.

1.2 User's Gender

The literature in nonverbal communication shows gender differences in processing nonverbal cues. Women pay more attention to nonverbal cues than men in general (Mazanec & McCall, 1976). Numerous studies also show that women are more accurate in decoding nonverbal cues (Hall, 1978). Furthermore, women are found to be more susceptible to the influence of nonverbal cues than men, for example, in inferring trait information (Zahn, 1973). The issue of face-voice pairing falls into the nonverbal domain because all face-voice combinations deliver the same verbal content. Thus, females' general tendency to be more sensitive to and concerned with nonverbal cues may apply to their reactions to face-voice pairing in talking heads. Therefore, the user's gender deserves consideration in examining users' responses to the different face-voice pairings of a talking head.

1.3 Context

The self-disclosure context is highly attitudinal, focusing on trust. The participants answered the questions on their own without being monitored and thus were not forced to disclose. All the disclosure questions were one-sentence long and written in simple English, and appear to have low cognitive demand. It is plausible that in a highly attitudinal context, users may be more alert and guarded and more likely to scrutinize the trustworthiness of the agent and thus more sensitive to the consistency issue.

One motivation for the present study is whether consistency preference would hold up in a context dominated by cognitive demand such as comprehension of long messages. In a comprehension-driven context, users may be more concerned with the clarity of the voice than the consistency between the face and the voice so that the human voice would prevail over the synthetic voice regardless of the face.

2 Method

To provide a full test of the consistency issue between human and synthetic faces and voices and investigate the roles of the user's gender and the nature of the context, a 2 (human face vs. synthetic face) by 2 (human voice vs. synthetic voice) by 2 (male vs. female users) by 2 (attitudinal vs. cognitive context) mixed-design experiment was conducted. The face, the voice, and the gender were between-participants

factors, while the context was the within-participants factor. The self-disclosure context was replicated. A cognitive context, where users listened to book reviews orally delivered by a talking head on a simulated book Web site, was added. The same talking head with one of the four combinations of human vs. synthetic faces and voices asked the participants the self-disclosure questions and presented the book reviews, respectively, in the two tasks. The order of the two tasks was balanced.

2.1 Participants

Eighty undergraduate students were recruited to participate in the study. Forty men and 40 women were randomly assigned to the four combinations of human vs. synthetic faces and voices. All participants were native English speakers to avoid potential language problem in processing the speech, especially the synthetic speech.

2.2 Procedure

The participants completed the study in a media laboratory. To reduce reactivity between the two experimental tasks, the two tasks were framed as *two* experiments which were conducted on two separate computers located on the opposite sides of the lab. After the participants finished one task on one computer, they received an instruction on the computer screen to approach the experimenter, who then led them to a second computer for the second task. This transition was intended as a psychological break for the participants between the two tasks.

The participants wore headphones during both experimental tasks. In the self-disclosure task, the talking head as the interviewing agent asked the participants four casual warm-up questions and nine intimate disclosure questions, same as in Gong et al.'s (2001) study. The self-disclosure questions were adapted from Moon (2000). Casual warm-up questions about hometown and hobbies were included because research has shown that disclosure escalates gradually in social interaction (Altman & Taylor, 1973) and human-computer interaction (Moon, 2000). Participants typed their answers in a text box on the screen and then clicked a "Submit" button to proceed to the next question. After all the questions, the participants filled out a questionnaire on the computer.

In the book-comprehension task, the same talking head was framed as a book-presenting agent and delivered five book reviews orally one at a time. The participants were told in the instruction that they could only listen to the book reviews once. The purpose was to stress the cognitive demand of the task. After they heard a book review, they clicked a "Continue" button to proceed to a new screen to answer a series of questions assessing their intent of consuming the book, evaluation of the book review, and perceived understanding of the book review. The first book was treated as a practice round. After going through the five book reviews, the participants filled out an attitudinal questionnaire on the computer.

After they finished both experimental tasks, the participants filled out a short paper-and-pencil exit questionnaire. The questionnaire contained items measuring their liking of the face and the voice and the manipulation checks for the experimental stimuli.

2.3 Manipulation

The synthetic face was the "Baldi" face provided in the CSLU Toolkit (http://cslu.cse.ogi.edu/toolkit). "Baldi" is a male and Caucasian facial model. The human face was a videotaped face of a male, Caucasian, and native English-speaking actor. Still pictures of the videos containing the synthetic face and the human face are displayed below (see Figure 1). The size of the video frames was 9.60 cm by 7.15 cm on a 38.1 cm-diagonal computer monitors.

Figure 1: Still pictures of the human face (left) and the synthetic face (right) (The originals were in color).

The synthetic voice was created using the Festival TTS engine in the CSLU Toolkit. The human voice was the recorded voice of the human actor. The synthetic face was automatically lip-synchronized with both synthetic speech and recorded human speech using the CSLU Toolkit. For the stimulus of the human face with the human voice, the face and voice of the human actor were simultaneously captured on the videotape. For the stimulus of the human face talking with the synthetic voice, the human actor was trained to talk over the synthetic voice in synchrony. The video files of the actor talking over the synthetic voice were then dubbed in synchrony with the sound files of the synthetic voice.

The researchers chose five fiction books that were unlikely to have been read by the participants. There were questions after each book regarding whether the participants had read the book or other book(s) by the author. The book reviews were modified from the ones in Amazon.com. The length of the book reviews ranged from 134 to 193 words. The length of the video files delivering the book reviews ranged from 47.5 to 72.2 seconds. The length of the video file for any of the book reviews was similar among the four versions of the talking head with the maximal difference being 2.9 seconds (4.20%).

2.4 Measures

Trust

For the self-disclosure task, the answers to the nine open-ended intimate disclosure questions were measured on two dimensions. The *amount of disclosure* was measured with a word count of the responses per question. Such answers as "I don't know", "I don't want to tell you", or "You are so nosy" counted as zero. The reliability of this index was high (α = .80). The *depth of self-disclosure* was based on intimacy ratings. Two coders, who were blind to the hypotheses and experimental conditions, independently rated each of the participants' responses on a five-point Likert-type scale (1 = "low intimacy", 5 = "high intimacy"). The inter-rater reliability was high (r = .83). Disagreements were resolved by averaging. The reliability of the index was acceptable (α = .78).

In addition to the behavioral measure of trust using self-disclosure, an attitudinal scale measured the participants' trust of the talking head in the questionnaire after the self-disclosure task. The scale was modified from the Individualized Trust Scale (Wheeless & Grotz, 1977), excluding the items "*exploitive-benevolent*" and "*faithful-unfaithful*" because they appeared not applicable to this scenario. After factor analysis, six items were selected to form the measure: *trustworthy-untrustworthy, distrustful of the agent-trustful of the agent, confidential-divulging, safe-dangerous, respectful-disrespectful,* and *unreliable-reliable.* Reliability was high (α = .87).

Effects in the Cognitive Task

In the book-comprehension task, the questions after each book review assessed three constructs:

C*onsumption intent about the book:* It was measured by three questions: *"How likely would you be to enjoy reading this book?", "How likely would you be to recommend this book to your friends?",* and *"How likely would you be to buy this book?"* on 10-point Likert-type scales ("1" = "very unlikely", "10" = "very likely"). The averaged responses to these questions across the four books formed this measure (α = .91).

Evaluation of the book review was measured by *"How would you rate the quality of the book review?"* (on a 10-point Likert scale, "1" = "very low quality", "10" = "very high quality") and *"How much did you like the book review?"* (on a 10-point Likert scale, "1" = "not at all", "10" = "very much"). Reliability of this measure was very high (α = .90).

Perceived understanding of the book review: measured by *"How well did you understand the book review?"* on a 10-point Likert scale ("1" = "very poorly", "10" = "very well").

Other Attitudes towards the Talking Head and User Experience

Except the trust scale, the questionnaires after both tasks contained similar attitudinal measures. The participants judged how well a series of adjectives described the interviewing agent or the book-presenting agent and how they felt during the tasks on 10-point Likert-type scales ("1" = "describes very poorly', "10" = "describes very well").

Negativity of the agent: consisted of *disconcerting, disturbing,* and *frustrating*. Reliability was high (disclosure task: α = .80; book task: α = .84).

Strangeness of the agent: consisted of *strange, surprising,* and *unusual* (α = .75 and .76, respectively).

Clarity of the agent: consisted of *articulate, clear,* and *hard-to-understand* (α = .80 and .88, respectively).

Rudeness of the interviewing agent (for the disclosure task only): consisted of *intrusive, offensive,* and *rude* (α = .84). This measure was included for the disclosure task because the participants were expected to have no prior exposure with the talking head and the disclosure questions were intimate.

Enjoyment of the user: consisted of *enjoyable, entertained, fun, happy,* and *pleasant* (α = .85 and .86, respectively).

Hesitance to disclose (for the disclosure task only): consisted of *guarded, hesitant,* and *questioning* (α = .76). This measure provided an additional assessment of the participants' trust.

The paper-and-pencil exit questionnaire contained measures about liking of the face and the voice of the talking head. Seven-point semantic differential scales were used.

Liking of the face: consisted of five pairs of adjectives: *pleasant-unpleasant, likeable-annoying, cold-warm, disturbing-pleasing,* and *nice-awful* (α = .81).

Liking of the voice: consisted of three pairs: *jarring-soothing, unpleasant-pleasant,* and *awful-nice* (α = .86).

Time-on-Task Measures

The server running the study recorded the time that each new computer screen in the process of the experiment loaded. Because inconsistency in the talking head may affect the participants' process of judging the talking head as in judging a human with inconsistent traits (Schul et al., 1983), the time taken in answering questions about the talking head was calculated. Time on answering questions about each book and book review was also calculated to discern potential second-order effect on judgment of the book and book review.

3 Results

Manipulation checks showed that the participants correctly recognized the human face and voice as "of a real person" and the synthetic face and voice as "computer-synthesized". They also perceived the human face with the human voice and the synthetic face with the synthetic voice as significantly more "consistent" and "matched" than the other two combinations. None of the participants reported that they had read any of the books or other books by the authors. None of the participants reported prior exposure to the human face or voice or the synthetic face. Four participants reported that they had heard the synthetic voice before but were not familiar with it. T-tests between the two orders of the experimental tasks showed the task order did not cause significant differences on the dependent measures.

3.1 Trust

Supporting the face-voice consistency argument, a significant face-voice interaction effect was found for the amount of self-disclosure, $F(1, 72) = 17.02$, $p < .001$. Both male and female participants disclosed more about themselves when the human face was paired with the human voice than with the synthetic voice and when the synthetic face was paired with the synthetic voice than with the human voice (Males: $F(1, 36) = 14.47$, $p < .001$; Females: $F(1, 36) = 5.48$, $p < .05$) (see Figure 2). There were no significant effects on the depth of disclosure in this study.

(For all the graphs, HF = human face, SF = synthetic face, HV = human voice, SV = synthetic voice)

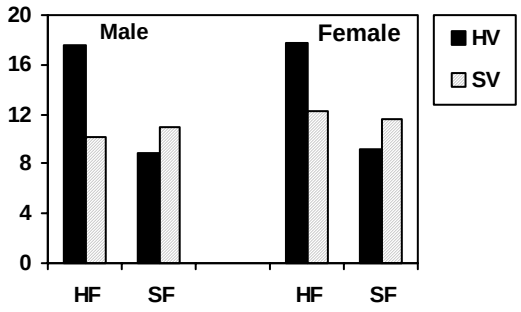

Figure 2. Amount of disclosure.

Additional evidence supporting the effect of face-voice consistency on trust was found on the attitudinal measures of trust and hesitance to disclose in the questionnaire after the self-disclosure task. A significant face-voice interaction effect was found for trust, $F(1, 72) = 13.27$, $p < .001$: both genders trusted the talking head more when it had face-voice consistency than inconsistency (Males: $F(1, 36) = 11.05$, $p < .01$; Females: $F(1, 36) = 4.45$, $p < .05$) (see Figure 3). Furthermore, females felt more hesitant to disclose when the face and the voice were inconsistent than consistent, $F(1, 36) = 6.60$, $p < .05$ [i]. The interaction effect was not significant for males, $F(1, 36) = 0.42$, $p > .52$. These results lend additional support to the argument that when the face and the voice of the talking head are consistent, people trust it more, are more willing to disclose about themselves, and do disclose more about themselves.

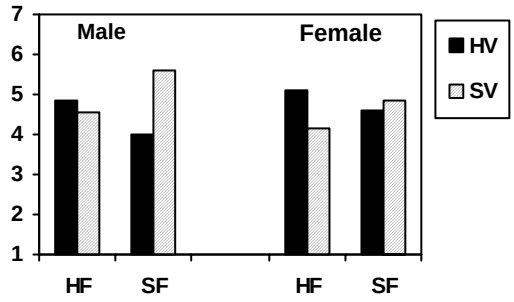

Figure 3. Trust of the talking head in the disclosure task.

3.2 Effects on the Cognitive Task

No significant face-voice interaction effects were found on the measures for the book-comprehension cognitive task. The main effects of the voice were the dominant finding. Both male and female participants had greater intent of consuming the books when the book reviews were delivered by the human voice than by the synthetic voice, regardless of the face (Males: $F(1, 36) = 6.09$, $p < .05$; Females: $F(1, 36) = 21.49$, $p < .001$) (see Figure 4). The same pattern of results emerged for the evaluation of the book reviews (Males: $F(1, 36) = 9.92$, $p < .01$; Females: $F(1, 36) = 45.23$, $p < .001$) and the perceived understanding of the book reviews (Males: $F(1, 36) = 5.75$, $p < .05$; Females: $F(1, 36) = 59.91$, $p < .001$).

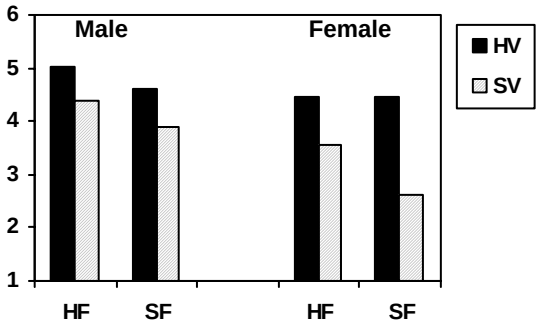

Figure 4. Consumption Intent about the books.

[i] Due to space limit, some graphs of the results are omitted.

3.3 Attitude towards the Talking Head and User Experience

For *negativity* of the talking head in the disclosure task, the face-voice interaction was not significant for males ($F(1, 36) = 0.11$, $p > .74$), but significant for females in support of consistency ($F(1, 36) = 17.69$, $p < .001$): Talking heads with inconsistent face-voice pairings were perceived more negatively than talking heads with consistent pairings. The same pattern of results emerged for the book task (Males: $F(1, 36) = 0.21$, $p > .65$; Females: $F(1, 36) = 26.90$, $p < .001$). The different reactions to face-voice pairings between male and female participants caused significant three-way interaction ($F(1, 72) = 7.93$, $p < .01$ in the disclosure task; $F(1, 72) = 14.31$, $p < .001$ in the book task).

For the *strangeness* of the talking head in the disclosure task, the face-voice interaction was not significant for males ($F(1, 36) = 0.02$, $p > .89$), but significant for females in support of consistency ($F(1, 36) = 11.23$, $p < .01$). The same pattern of results was found in the book task (Males: $F(1, 36) = 0.01$, $p > .91$; Females: $F(1, 36) = 19.64$, $p < .001$). These different reactions to face-voice pairings between male and female participants caused significant three-way interaction ($F(1, 72) = 6.03$, $p < .05$ in the disclosure task; $F(1, 72) = 6.53$, $p < .01$ in the book task). Females thought the talking head was less *rude* when it had consistent face-voice pairings than inconsistent pairings ($F(1, 36) = 11.47$, $p < .01$), while males did not show significant preference ($F(1, 36) = 0.02$, $p > .90$). This gender difference caused a significant three-way interaction, $F(1, 72) = 6.41$, $p < .01$.

For the *clarity* of the talking head, no significant face-voice interaction effects were found. Instead, the human voice prevailed over the synthetic voice regardless of the face in both tasks among both genders (Disclosure task: males: $F(1, 36) = 7.01$, $p < .01$; females: $F(1, 36) = 39.62$, $p < .001$; Book task: males: $F(1, 36) = 15.04$, $p < .001$; females: $F(1, 36) = 91.88$, $p < .001$).

For the *enjoyment* of the users in the disclosure task, significant face-voice interaction among males showed that they enjoyed the task more when the talking head had consistent face-voice pairings than inconsistent pairings, $F(1, 36) = 5.95$, $p < .05$. The interaction was not significant for females, $F(1, 36) = 1.69$, $p > .20$. This gender difference caused a significant three-way interaction effect ($F(1, 72) = 7.49$, $p < .01$). Overall, males found the task more enjoyable than females, $F(1, 72) = 15.83$, $p < .001$. The same pattern of results emerged in the book task in terms of the face-voice interaction (Males: $F(1, 36) = 4.11$, $p < .05$; Females: $F(1, 36) = 0.01$, $p > .92$).

For *liking of the face* measured in the exit questionnaire, a significant face-voice interaction among females showed that the face was liked more when it was paired with the consistent voice, $F(1, 36) = 18.88$, $p < .001$. The interaction was not significant for males, $F(1, 36) = 0.13$, $p > .73$. Overall, males liked the face more than females, $F(1, 72) = 21.43$, $p < .001$. Because of the gender difference, there was a significant three-way interaction, $F(1, 72) = 6.96$, $p < .01$. For *liking of the voice*, not surprisingly, the human voice was liked more than the synthetic voice regardless of the face for both genders (Male: $F(1, 36) = 26.00$, $p < .001$; Females: $F(1, 36) = 24.53$, $p < .001$). Interestingly, a face-voice consistency interaction effect also emerged for females, who liked the voice more when it accompanied the consistent face, $F(1, 36) = 9.37$, $p < .01$. Males overall liked the voice more than females, $F(1, 72) = 12.35$, $p < .001$.

Thus, it appears that females showed more consistency preference in their attitudinal responses than males. A SIGN statistical test supported this observation, $p < .05$.

For the attitudinal measures which were identical in the two tasks, repeat-measures ANOVA's were conducted to discern any possible difference due to the task factor. There were no significant differences between the two tasks except for the clarity of the talking head in females' responses: Females' preference of the human voice was stronger in the book task than in the disclosure task, $F(1, 36) = 8.93$, $p < .01$. This makes sense given the comprehension nature of the book task.

3.4 Time Measures

Significant face-voice interaction effects were found for the time on the questions judging the talking-head agent in both tasks for both genders (Disclosure task: males: $F(1, 36) = 6.58$, $p < .05$; females: $F(1, 36) = 9.82$, $p < .01$, see Figure 5; Book task: males: $F(1, 36) = 7.04$, $p < .01$, females: $F(1, 36) = 10.16$, $p < .01$). In line with the literature on inconsistency (Schul et al., 1983), participants spent longer time judging the talking head when it had inconsistent face-voice pairings than consistent pairings. As a second-order effect, male participants spent significantly longer time forming opinions about the books and book reviews when the talking head had inconsistent face-voice pairings, $F(1, 36) = 5.70$, $p < .05$. The same pattern was observed among females, approaching significance statistically, $F(1, 36) = 2.23$, $p > .14$.

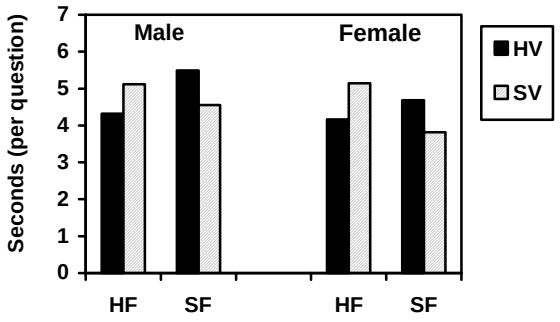

Figure 5. User's time spent judging the talking head in the disclosure task.

3.5 Summary

Significant face-voice interaction effects in support of consistency were found on the amount of disclosure, the trustworthiness of the talking head in the disclosure task, and time on judging the talking head in both tasks for both genders. For females, consistency effect was also found on hesitance to disclose and rudeness of the talking head in the disclosure task, negativity and strangeness of the talking head in both tasks, and liking of the face and the voice. For males, consistency effect was also found on enjoyment of both tasks and time on judging the books and book reviews.

Main effects of the voice showed the superiority of the human voice for consumption intent about the books, evaluation of the book reviews, perceived understanding of the reviews, clarity of the talking head in both tasks, and liking of the voice for both genders.

4 Discussion

The study demonstrates that face-voice consistency in a talking head is desirable, achieving more positive effects on users' trust in terms of trust-driven behavior and attitudinal responses of trust. Face-voice consistency also achieved more positive responses in terms of other attitudes from users, particularly from female users compared to male users, in both the attitudinal and the cognitive contexts. Face-voice consistency also appear to require less processing time compared to face-voice inconsistency. The results are in line with the psychological literature that shows consistency as a principle in human perception and social interaction.

Another significant finding is that context of human-computer interaction proved important. Voice clarity became the dominant factor in a comprehension-driven cognitive context and achieved more positive task outcomes, overwhelming face-voice consistency. Thus, contextual specificity seems to modify consistency as a general principle.

In addition, users' gender emerged as an important factor. Female users showed stronger attitudinal preference for face-voice consistency and more negative attitude towards the synthetic face and voice. This can be explained by females' general tendency of being more sensitive to and concerned with nonverbal cues. They may care more about face-voice consistency and prefer more the natural human face and voice than males.

4.1 Limitations and Future Research

The study is limited in terms of the gender of the talking head, context, and human-likeness level of the synthetic face and voice. It points out important directions for future research.

Because the talking heads in the study were male, it is unknown whether the found gender differences were due to the females' general tendency in processing nonverbal cues or due to the male identity of the talking head. Thus, future research should include a female talking head.

Entertainment context is an important type of context worth examining in future research. Animated characters in films commonly talk with human voices, counter to the consistency claim. In an entertainment context, users generally have low personal stake with low potential loss or worry, compared to the privacy concern in trust-laden contexts such as self-disclosure. According to the Elaboration Likelihood Model (Petty & Cacioppo, 1986), people's processing in a context with low personal stake tends to fall in the peripheral route of processing with less scrutiny, so that the face-voice consistency issue may be not as salient as in a context causing central-route processing and high scrutiny from users.

As synthesis of faces and speech keeps advancing, they are starting to differ substantially in terms of how human-like they appear. An interesting question, then, is whether a very human-like synthetic face or voice should be paired with a synthetic counterpart or a human counterpart in comparison to a much less human-like face or voice. This will also help us understand how users perceive and distinguish the human versus humanoid category and instantiations.

In addition to pairing face and voice for a talking head, consistency seems an important issue for a variety of topics in interfaces and human-computer interaction because consistency is a general principle in human perception and interactions. Valuable topics for testing consistency include relationships between emotional expression of an agent and the emotion tone of the interface context, between attractiveness of an agent and its behavioral intelligence and performance, between characteristics of an agent and those of the user in terms of age, gender, ethnicity, cultural background, personality, and so on, between an agent

and the party being represented by the agent in computer-mediated communication and collaboration, and between an agent and interface and the culture they represent or target.

References

Altman, I., & Taylor, D. A. (1973). *Social penetration: The development of interpersonal relationships*. New York, NY: Hold, Rinehard and Winston.

Argyle, M., Alkema, F., & Gilmour, R. (1971). The communication of friendly and hostile attitudes by verbal and non-verbal signals. *European Journal of Social Psychology, 1*, 385-402.

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

Ekman, P., & Friesen, W. V. (1969). Nonverbal leakage and clues to deception. *Psychiatry, 32*, 88-95.

Gong, L., Nass, C., Simard, C., & Takhteyev, Y. (2001). When non-human is better than semi-human: Consistency in speech interfaces. In M. J. Smith & G. Salvendy & D. Harris & R. Koubek (Eds.), *Usability evaluation and interface design: Cognitive engineering, intelligent agents, and virtual reality* (pp. 1558-1562). Mahwah, NJ: Lawrence Erlbaum Associates.

Graves, J. R., & Robinson, J. D. (1976). Proxemic behavior as a function of inconsistent verbal and nonverbal messages. *Journal of Counseling Psychology, 23*, 333-338.

Hall, J. A. (1978). Gender effects in decoding nonverbal cues. *Psychological Bulletin, 85*, 845-857.

Hendrick, C. (1972). Effects of salience of stimulus inconsistency on impression formation. *Journal of Personality & Social Psychology, 22*, 219-222.

Isbister, K., & 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.

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

Lundeberg, M., & Beskow, J. (1999). Developing a 3D-agent for the August dialogue system, *Proceedings of AVSP'99* (pp. 151-156).

Massaro, D. M. (1998). *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*. Cambridge, MA: MIT Press.

Mazanec, N., & McCall, G. J. (1976). Sex factors and allocation of attention in observing persons. *Journal of psychology, 93*, 175-180.

Moon, Y. (2000). Intimate exchanges: Using computers to elicit self-disclosure from consumers. *Journal of Consumer Research, 26*, 323-339.

Nass, C., & 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.

Petty, R. E., & Cacioppo, J. T. (1986). *Communication and persuasion: Central and peripheral routes to attitude change*. New York, NY: Springer/Verlag.

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

Schul, Y., Burnstein, E., & Martinez, J. (1983). The informational basis of social judgments: Under what conditions are inconsistent trait descriptions processed as easily as consistent ones? *European Journal of Social Psychology, 13*, 143-151.

Wheeless, L. R., & Grotz, J. (1977). The measurement of trust and its relationship to self-disclosure. *Human Communication Research, 3*, 250-257.

Zahn, G. L. (1973). Cognitive integration of verbal and vocal information in spoken sentences. *Journal of Experimental Social Psychology, 9*, 320-334.