Can Users Live with Overconfident or Unconfident Systems? : A Comparison of Artificial Subtle Expressions with Human-like Expression

Takanori Komatsu
Shinshu University
3-15-1 Tokida, Ueda,
Nagano 386-8567, Japan.
tkomat@acm.org

Kotaro Funakoshi
Honda Research Institute Japan
8-1 Honcho, Wako,
Saitama 351-0188, Japan.
funakoshi@jp.honda-ri.com

Kazuki Kobayashi
Shinshu University
4-7-1 Wakasato, Nagano,
Nagano 380-8553, Japan.
kby@shinshu-u.ac.jp

Mikio Nakano
Honda Research Institute Japan
8-1 Honcho, Wako,
Saitama 351-0188, Japan.
nakano@jp.honda-ri.com

Seiji Yamada
National Institute of Informatics
SOKENDAI
2-1-2 Hitotsubashi, Chiyoda,
Tokyo 101-8430, Japan.
seiji@nii.ac.jp

CHI'12, May 5–10, 2012, Austin, Texas, USA.
ACM 978-1-4503-1016-1/12/05.

Abstract

We assume that expressing the levels of confidence using human-like expressions will cause users to have a poorer impression of a system than if artificial subtle expressions (ASEs) were used when the quality of the presented information does not match the expressed level of confidence. We confirmed that this assumption was correct by conducting a psychological experiment.

Keywords

Artificial Subtle Expressions (ASEs); Human-like Expressions; Confidence; Users' Subjective Impressions

ACM Classification Keywords

H.5.2 [User Interfaces]: Evaluation/methodology; J.4 [Social behavioral sciences]: Psychology.

General Terms

Experimentation, Human Factors

Introduction

Human-machine communication using speech sounds is becoming more common [1,2] because users can obtain information while engaging in their primary tasks without facing nor manually operating the information providing systems (e.g., intelligent home appliances or car navigation systems). However, due to various reasons [3], such as noise in the sensors, the incompleteness of data, immaturity of technology, and the complexity of tasks, the reliability of such systems is often limited. Cai et al. [4] experimentally showed how expressing the levels of confidence for such systems to indicate whether the system's represented information is accurate or not to users plays an important role in improving both the user's performance and their impressions.

When intending to express a system's level of confidence, one can easily have the idea of using human-like verbal expressions such as "probably," "definitely," or "83% confident." However, expressing levels of confidence using such human-like expressions might frustrate users when the quality of the presented information does not match the expressed level of confidence. For example, users might feel frustrated with systems (like car navigation systems) that express a higher level of confidence like "you should follow my suggested route" or "I am 80% confident," but the represented information was wrong (this is the case of being "overconfident"). Since human-like expressions make users expect higher human-like abilities from the systems [5], such inconsistent behaviors eventually make them deeply disappointed [6,7].

Related to the above issue, we have proposed artificial subtle expressions (ASEs) as an intuitive methodology for notifying users of a system's internal state. In particular, we showed that ASEs implemented as beep-like sounds succeeded in accurately and intuitively conveying a system's confidence to the users [8,9]. Therefore, we assume that our proposed ASEs are suitable for expressing levels of confidence in comparison to human-like expressions.

The purpose of this study is then to confirm the above assumption that expressing levels of confidence using human-like expressions gives users a poorer impression of a system than by using those expressed by ASEs when the quality of the presented information does not match the expressed levels of confidence (in particular, where the system's suggestions are incorrect/correct even though the expressed confidence is high/low) by conducting a psychological experiment to comprehend the users' subjective impressions. Such inconsistency between the represented information and the level of confidence is inevitable due to the immaturity of the current technology used in media terminals and due to the fact that the levels of confidence are just a probability indicating how accurate the represented information is. Therefore, this study should contribute to proposing a novel interaction technique on how to handle this inconsistency without frustrating the users.

Experiment

Environment

We used a "driving and treasure hunting" video game as our experimental environment for comprehending the participants' impressions of a system. In this game, a game image scrolls forward on a straight road as if the participant is driving a car using a navigation system with three small mounds of dirt appearing along the way. A coin is inside one of the three mounds, while the other two mounds contain nothing. The game ends after the participant encounters 20 sets of mounds (20 trials). The purpose of this game is to get as many coins as possible. The location of the coin amongst the three mounds was randomly assigned. In each trial, the navigation system next to the driver's seat (circle on left of Fig. 1) told them which mound it expects the coin to be in by using speech. The participants could freely accept or reject the navigation system's suggestions. After the participant selected one mound among the three using a computer mouse, they could immediately know whether the selected mound contained the coin or not on the display (middle of Fig. 1).

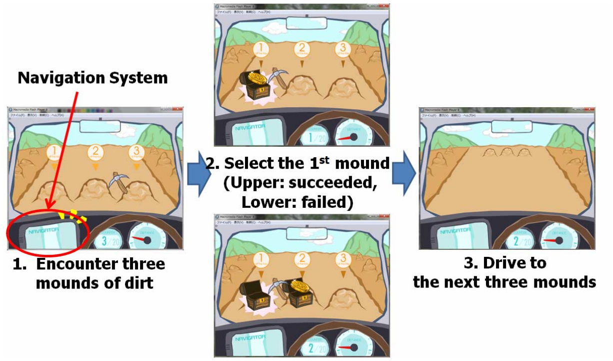
Navigation System
1. Encounter three mounds of dirt
2. Select the 1st mound (Upper: succeeded, Lower: failed)
3. Drive to the next three mounds

Figure 1. Driving and treasure hunting video game

Using Speech Sounds

In this experiment, the navigation system used Japanese speech sounds to suggest to the users the expected location of the coin; that is, "ichi-ban (no. 1)," "ni-ban (no. 2)," or "san-ban (no. 3)." These speech sounds were created by adding robotic-voice effects to the recorded speech sounds of one of the authors. These sounds were the main protocol (suggestion) of the navigation system. We then prepared the following three experimental stimuli (conditions) to express the levels of confidence of the main protocol.

- **ASE Condition**: One of the two ASEs was played 0.2 seconds after the speech sounds. These two ASEs were triangular wave sounds 0.5 seconds in duration with different pitch contours; that is, one was a flat ASE (onset F0: 250 Hz and end F0: 250 Hz) and the other was a decreasing ASE (onset F0: 250 Hz and end F0: 100 Hz). The suggestions with decreasing ASEs were able to inform users of the system's lower level of confidence in its suggestions while the ones with flat ASEs were to inform them of a higher level of confidence [8,9].
- **Paralinguistic Condition**: We prepared two stimuli for typical human-like expressions by modifying the paralinguistic information of the suggestions ("ichi-ban," "ni-ban," and "san-ban"), e.g., the rate of the utterances and intonation patterns; that is, one was an utterance with a faster rate with a falling intonation ("ichiban!"), while the other was a slower-rate utterance with a rising intonation ("i..chi..ba..n?"). We designed the latter stimulus (slower rate with rising intonation) to inform users of the system's lower level of confidence in the form of a question, while the former stimulus (faster rate and falling intonation) informs them of a higher level of confidence.
- **Linguistic condition**: We prepared two stimuli as another kind of typical human-like expression by adding Japanese linguistic suffixes to the suggestions; that is, one with "desu (definitely)" 0.1 seconds after the suggestion, and the other with "dato omoi masu (I guess so)." We designed the suggestions with "dato omoi masu" to inform users of the system's lower level of confidence, while the ones with "desu" to inform them of a higher level of confidence.

1. My feeling is the as usual even if I use this system.
2. This system has good adaptability.
3. This system deserves to work on responsible tasks.
4. I place complete reliance on this system.
5. This system makes favorable impressions on many people.
6. This system is always preferred among the similar systems.
7. I prefer this system because this system is similar to my way of thinking.
8. This system is human-like.
9. This system can offer good services.
10. I am satisfied with the services of this system.
11. I want to use this system again.
12. I cannot stand the mistakes made by this system.
13. This system is polite.
14. This system is a sufficiently reliable one.
15. This system is helpful for me.
16. This system is lovable.
17. I enjoy spending time with this system.
18. I feel tired when I use this system.

table 1. Utilized Questionnaire

Among the 20 trials, the navigation system expressed the information with a higher level of confidence 10 times and with a lower one 10 times. The order of these two levels of confidence was counterbalanced across the participants.

Participants

Twenty Japanese university students (15 men and 5 women; 21 - 28 years old) participated. They were randomly divided into the following two experimental groups in terms of the accuracy of the navigation system's levels of confidence.

- **Consistent Group** (10 participants): The participants in this group interacted with a system that expressed levels of confidence that were consistent with the correctness of the information it presented; that is, when the system expressed the information at a higher level of confidence, the rate of the suggested mound containing the coin was 100%, and when the system expressed the information with a lower level of confidence, the rate was 0%.
- **Inconsistent Group** (10 participants): The participants in this group interacted with a system that expressed levels of confidence that were inconsistent with the correctness of the information it presented; that is, when the system expressed the information with a higher level of confidence, the rate of the suggested mound containing the coin was 50%, and when the system expressed the information with a lower level of confidence, the rate was also 50%.

All the participants experienced all three experimental stimuli, so the experimental design was a 2 × 3 mixed design; that is, the between-factor was consistent (consistent/inconsistent groups), while the within-factor was the type of stimuli (ASEs/Paralinguistic/Linguistic conditions).

Procedure

We used a web-based questionnaire system to comprehend the participants' impression of the navigation systems and their performances and behaviors using the treasure hunting video game in this experiment. First, the system displayed a consent form and the instructions for the experiment. Before starting the treasure hunting game, the participants were asked to listen to test sounds via a speaker or headphones and to adjust the sound volume to a comfortable level. Afterwards, they played the treasure hunting video game three times to experience all three conditions. The order of these conditions was counterbalanced among the participants.

After finishing each condition, the participants were asked to fill in a questionnaire on the navigation system, which consisted of 18 questions using a 7-point Likert scale (maximum evaluation: 7 points; minimum evaluation: 1 point). The summed points of these questions were used as the "participants' subjective impression scores" of this navigation system; that is, more points meant a better impression of the system (the highest score was 126 points and the lowest was 18). The questionnaire consisted of a modified love-liking scale [10] and our original questions (Table 1: Cronbach's alpha: 0.86).

Assumption

We assumed that expressing the levels of confidence using human-like expressions would give the users a poorer impression of the system than expressing these levels with ASEs when the quality of the presented information does not match the expressed levels of confidence; in particular, in the case where the systems' suggestions were incorrect/correct even though the confidence was high/low. That is, if we could observe that the participant's impression scores for the ASE condition were significantly higher than those for the paralinguistic and linguistic conditions in the inconsistent group, we would be able to verify our assumption.

Results

As a manipulation check, we confirmed that the type of experimental stimuli did not affect the participants' performance and behavior in this game. Therefore, we focused purely on the effects the types of experimental stimuli had on the participants' subjective impressions scores. The users' subjective impression scores for each group and condition are shown in Fig. 2. For the 10 participants in the consistent group, the average impression score for the ASE conditions was 58.2 (SD = 15.53), that for the paralinguistic ones was 75.1 (SD = 9.27), and that for the linguistic ones was 68.1 (SD = 68.1). For the 10 participants in the inconsistent group, the average impression score for the ASE conditions was 60.0 (SD = 11.40), that for the paralinguistic ones was 49.4 (SD = 11.94), and that for the linguistic ones was 48.6 (SD = 12.31).

These subjective impression scores were then analyzed using a 2 × 3 mixed ANOVA (between independent variable: consistent/inconsistent group, within independent variable: ASEs/paralinguistic/linguistic, and dependent variable: users' subjective impression scores). The results of the ANOVA showed significant differences in the interaction effect [F(2,36)=11.50, p<.01(**), effect size: η^2=0.15] and the main effect between independent variables [F(1,18)=9.99, p<.01(**), η^2=0.22]. The simple main effects of the between and within independent variables were then analyzed, and the results showed significant differences in the scores for the paralinguistic and linguistic conditions between the consistent and inconsistent groups [paralinguistic: F(1,18)=26.03, p<.01(**), linguistic: F(1,18)=11.71, p<.01(**)], and in the scores for the three experimental stimuli within both groups [consistent: F(2,36)=7.97, p<.01(**), inconsistent: F(2,36)=4.48, p<.05(*)]. A multiple comparison using an LSD test on the simple main effect of the within independent variables showed that the scores for the paralinguistic and linguistic conditions in the consistent group were significantly higher than those for the ASEs (MSe=90.4883, 5% level). In comparison, the scores for the paralinguistic and linguistic conditions in the inconsistent group were significantly lower than those for the ASEs (MSe=90.4883, 5% level).

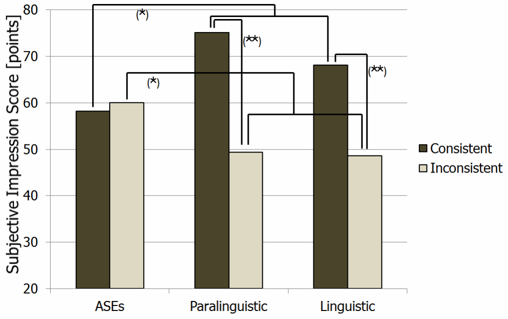

Figure 2. Subjective impression scores for each experimental condition and group

Therefore, we clearly observed that the users' impression scores for the ASE conditions were significantly higher than those for the paralinguistic or linguistic conditions in the inconsistent group, so we were able to verify our assumption. Moreover, we found that the scores for the ASE conditions in both groups were almost the same, while the scores for the paralinguistic and linguistic conditions significantly differed. Therefore, this also implies that the users' subjective impressions for a system expressing ASEs were quite robust regardless of the consistency between the represented information and the levels of confidence.

Conclusion

The results of this study showed that expressing confidence using human-like expressions received a higher evaluation in comparison to using such expressions with ASEs when the accuracy of the system's levels of confidence was perfect (in consistent group). However, it is almost impossible to build a user interface that can always provide correct suggestions and levels of confidence to users based on the level of current technology. Therefore, establishing a concrete methodology for handling the inconsistency between the suggestion and the levels of confidence is indispensable and worthwhile for the HCI domains.

We now have to investigate whether the acquired results can be used in much more realistic applications, e.g., spoken dialogue systems like an actual car navigation system. If we succeed, we can strongly argue that expressing the levels of confidence of systems by using ASEs is a reasonable methodology for avoiding frustrating users, and will contribute to the proposals for a novel interaction technique for dealing with the inconsistency between the represented information and the levels of confidence.

References

[1] Nass, C. and Brave, S. Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship, The MIT Press, MA, USA, 2005.

[2] Cohen, M. H., Giangola, J. P., and Balogh, J. Voice User Interface Design, Addison-Wesley, MA, USA, 2004.

[3] Benzeghibaa, M., De Moria, R., Derooa, O., Dupont S., Erbesa, T., Jouveta, D., Fissorea, F., Lafacea, P., Mertinsa, A., Risa, C., Rosea, R., Tyagia, V., and Wellekensa, C. Automatic speech recognition and speech variability: A review, Speech Communication 49, 10-11 (2007), 763-786.

[4] Cai, H. and Lin, Y. Tuning Trust Using Cognitive Cues for Better Human-Machine Collaboration, In Proc. HFES2010, pp. 2437-2441(5).

[5] Kanda, T., Miyashita, T., Osada, T., Haikawa, Y., and Ishiguro H. Analysis of Humanoid Appearances in Human-robot Interaction, IEEE Transactions on Robotics 24, 3 (2008), 725-735.

[6] Aronson, E. and Linder, D. Gain and loss of esteem as determinants of interpersonal attractiveness, Journal of Experimental Social Psychology 1, (2) 1965, 156-171.

[7] Komatsu, T., Kurosawa, R., and Yamada, S. How Does the Difference Between Users' Expectations and Perceptions About a Robotic Agent Affect Their Behavior?, International Journal of Social Robotics, DOI=10.1007/s12369-011-0122-y.

[8] Komatsu, T., Yamada, S., Kobayashi, K., Funakoshi, K., and Nakano, M. Artificial Subtle Expressions: Intuitive Notification Methodology of Artifacts, In Proc. CHI2010, ACM Press (2010), 1941-1944.

[9] Funakoshi, K., Kobayashi, K., Nakano, M., Yamada, S., and Komatsu, T. Non- humanlike Spoken Dialogue: a Design Perspective, In Proc. SIGDIAL2010 (2010), 176-184.

[10] Rubin, Z. Measurement of romantic love, Journal of Personality and Social Psychology 16, (2) 19P70, 265-273