Response Times when Interpreting Artificial Subtle Expressions are Shorter than with Human-like Speech Sounds

Takanori Komatsu
FMS, Meiji University
4-21-1 Nakano, Tokyo, Japan
tkomat@meiji.ac.jp

Kazuki Kobayashi
Shinshu University
4-17-1 Wakasato, Nagano, Japan
kby@cs.shinshu-u.ac.jp

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

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

Mikio Nanano
Honda Research Institute Japan Co., Ltd.
8-1 Honcho, Wako, Saitama
nakano@jp.honda-ri.com

ABSTRACT
Artificial subtle expressions (ASEs) are machine-like expressions used to convey a system's confidence level to users intuitively. In this paper, we focus on the cognitive loads of users in interpreting ASEs in this study. Specifically, we assume that a shorter response time indicates less cognitive load, and we hypothesize that users will show a shorter response time when interpreting ASEs compared with speech sounds. We succeeded in verifying our hypothesis in a web-based investigation done to comprehend participants' cognitive loads by measuring their response times in interpreting ASEs and speeches.

Author Keywords
Artificial Subtle Expressions (ASEs); Response time; Speech; Cognitive loads.

ACM Classification Keywords
H.5.2. User Interfaces: Evaluation/methodology; J.4. Social and behavioral sciences: Psychology.

INTRODUCTION
Speech interface systems, such as Apple's Siri or OK Google, are already popular and installed in most smartphones. However, there is still little possibility that speech interface systems will ever be 100% reliable [3,18,21,22]. To manage errors or misunderstanding caused by such systems, some studies have been focusing on displaying a system's confidence level to users, and these studies have shown that it is actually effective for various aspects of interaction between humans and systems [4,6,7,9,10]. For example, Antifakos et al. [1] showed that users adapt to a system easily if the system's confidence is displayed on a computer screen. Therefore, expressing a system's confidence to users is now becoming indispensable for speech interface systems. Most of these studies used speech sounds as human-like expressions to express their confidence level to users.

CHI 2017, May 06-11, 2017, Denver, CO, USA
 ISBN 978-1-4503-4655-9/17/05…$15.00
DOI: http://dx.doi.org/10.1145/3025453.3025649

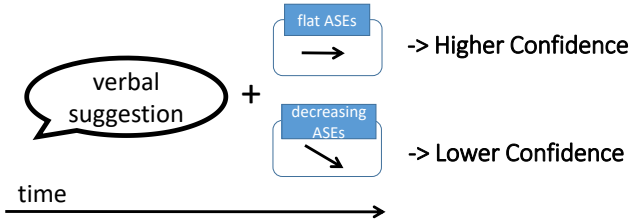

Figure 1. Artificial subtle expressions (ASEs)

In contrast with the above approaches, Komatsu et al. [14] proposed using artificial subtle expressions (ASEs) as machine-like expressions used to convey a system's confidence level to users intuitively. Although ASEs are similar to Earcons [5] at a glance, there are apparent differences between them [14]; that is, an Earcon is a brief sound that has an arbitrary mapping to a certain meaning, while an ASE is also a brief sound that has an inevitable mapping to a meaning based on psychological findings. Moreover, Earcons play a main role in the communication protocol, while ASEs play a complementary one. Specifically, they proposed two simple beeping like sounds used as ASEs: a flat sound (flat ASE) and a sound with a decreasing pitch (decreasing ASE). These ASEs were added after the system's verbal suggestions. They then showed that suggestions made with decreasing ASEs conveyed a low system confidence level to users intuitively (Figure 1). Up to now, several studies on ASEs have reported on the advantages of ASEs, that is, being suitable for imperfect systems [15] and language-independent interpretations being possible [16].

In this study, we focus on users' cognitive loads consumed in interpreting ASEs. Specifically, we hypothesize that users require less cognitive load when interpreting ASEs

compared with speech sounds used as human-like expressions. While there are several methodologies for comprehending users' cognitive loads, such as the dual task paradigm [12,19,24] or measuring users' biological signals [2,8], we focus on the response time for interpreting ASEs as an objective indicator of users' performance [11,23]; that is, a shorter response time indicates less cognitive load. Although the reaction time itself may not be enough to comprehend the exact amount of participants' cognitive loads, it can be utilized to relatively comprehend whether ASEs or speech require a lower cognitive load. According to the above hypothesis, we assume that users will show a shorter response time when interpreting ASEs compared with speech sounds. We thus conducted a web-based experiment to measure users' response times for both ASEs and speech sounds and to investigate whether the above hypothesis is true or not.

EXPERIMENT

Settings

We conducted a web-based experiment to measure participants' response time for ASEs and speech. We used a "driving treasure hunting" video game as an experimental environment (Figure 2). In this game, the game image scrolls forward on a straight road as if the participant is driving a car with a navigation system and with small three 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 participants encounter 24 sets of mounds (24 trials).

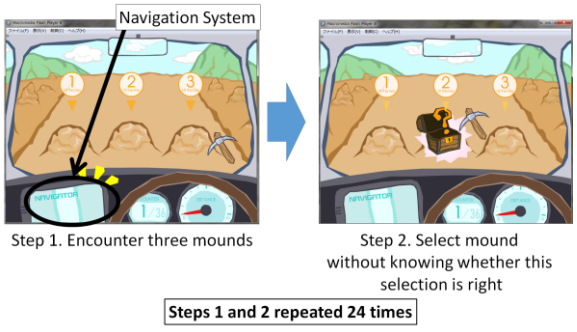

Figure 2. Driving treasure hunting video game

The purpose for the participants is to get as many coins as possible. The location of the coin among the three mounds is randomly assigned. In each trial, the navigation system to the left of the driver seat (circled in the left image of Figure 2) told them which mound it expected the coin to be in by using verbal suggestion with ASEs or speech. The participants could freely accept or reject the navigation system's suggestions. In each trial, even after the participants selected one mound among the three, they were not told whether the selected mound had the coin or not (only a question mark appearing from the opened treasure box is displayed, as shown in the right image of Figure 2). Here, if the participants received feedback on whether their selection was correct or not, they usually started to solve a three-armed bandit problem [13] by considering strategies like "after estimating the probability distribution on which mound has a coin in the initial trials, select the mound with the highest probability." We then avoided such strategic behavior by providing no feedback. The participants were then informed of their total numbers of coins only after they finished all 24 trials. It took about three minutes to complete this game.

Stimuli

In this experiment, the navigation system used Japanese speech to suggest to the participants 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 of one of the authors. The duration of these three sounds was 0.5 seconds. We then prepared the following two types of experimental stimuli to express levels of confidence to the participants.

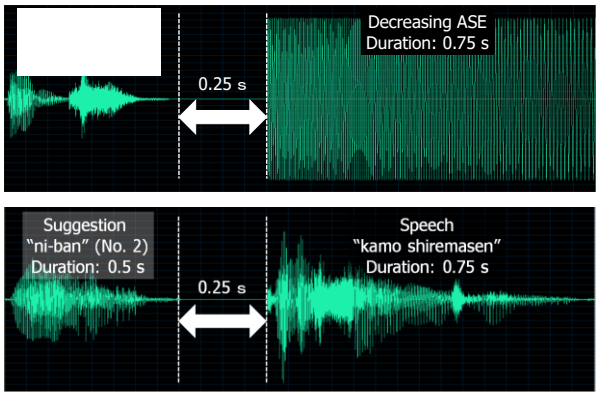

Figure 3. Speech waveform of suggestion "ichi-ban" with decreasing ASE (upper) and suggestion "ni-ban" with speech "kamo shiremasen" (lower)

- ASEs: One of two ASEs was played 0.25 seconds after the verbal suggestion (upper image of Figure 3). These two ASEs were triangular wave sounds 0.75 seconds in duration with different pitch contours; that is, one was a flat ASE (onset F0: 400 Hz and end F0: 400 Hz), and the other was a decreasing ASE (onset F0: 400 Hz and end F0: 250 Hz). The suggestions with decreasing ASEs were 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.
- Speech: As a kind of typical human-like expression, we prepared two stimuli by adding Japanese speech to the suggestions, that is, one with "dato omoi masu (I bet)" 0.25 seconds after the suggestion and the other with "kamo shiremasen (I am not sure)" (the lower image of Figure 3). We designed the suggestions with "kamo shiremasen" to inform users of the system's lower level of confidence, while the ones with "dato omoi masu" were designed to inform them of a higher level of confidence. The duration of both pieces of speech were

0.75 seconds, the same as the ASEs, and they were created by adding robotic-voice effects to the recorded speech sounds of one of the authors.

Here, the length of the suggestions with ASEs and with speech was completely the same at 1.5 seconds (suggestions: 0.5 seconds, interval: 0.25 seconds, and ASEs or speech: 0.75 seconds), and there were six variations for each type of stimulus (3 suggestions × 2 ASEs/pieces of speech).

Participants

163 Japanese undergrads participated. These participants voluntary responded to a call for participants from the authors. They were randomly assigned one of the following two experimental conditions.

- ASE condition (82 participants): Among 24 trials of the driving treasure hunting video game, 6 stimuli with ASEs (3 suggestions × 2 ASEs) were randomly presented to the participants 4 times.
- Speech condition (81 participants): Among 24 trials, 6 stimuli with speech (3 suggestions × 2 pieces of speech) were also randomly presented 4 times.

The URL of the web-based experiment system was given to these participants, and they were asked to open it on their own PCs. The system first displayed a consent form and instructions for the experiment. These instructions never mentioned or explained the ASEs or speech given after suggestions made to the participants, but they strongly mentioned that the participants should select the correct mound as fast as possible. Before starting the game, the participants were asked to listen to a test sound via speakers or headphones and to adjust the sound volume to a comfortable level. Afterward, they played the game. The response time was defined as the duration from the onset of the sound stimuli to the participant's mound selection, and this was automatically measured by the experimental system. The game was implemented with Adobe Flash, and the participants' response times were measured by the internal timer of their PCs. The game started after downloading all required data for the game, so the transmission speed between the experimental system and the participants' PCs did not affect the measurements of their response times. In terms of the adequacy of measuring response time in a web-based experiment, Komarov et al. [17] already reported that "there were no significant differences between the two settings (lab experiment and MTurk experiment) in the raw task completion times, error rates," so we assumed that this web-based experimental system was reasonable for measuring the participants' reaction times when interpreting ASEs and speech.

On the basis of the measured response time, we detected "lazy participants" who were not really listening to the given stimuli. Specifically, we eliminated the data of participants who took more than 10 seconds to respond at least one time (did not concentrate during the experiment), who responded within 0.05 seconds at least one time (did not listen to the given stimuli), and whose average response time was less than 0.5 seconds (did not listen to the ASEs or speech parts). As a result, the data for the 23 participants (12 participants for the ASE condition and 11 participants for the speech condition) were eliminated, so of the data for the remaining 140 participants (99 male, 39 female, and 2 unanswered, 20 – 23 years old), 70 of those were in the ASE condition, and the remaining 70 were in the speech condition.

Manipulation Check

The suggestions with a flat ASE and the phrase "dato omoi masu (I bet)" were expected to convey a higher confidence level to the participants, while those with a decreasing ASE and the phrase "kamo shiremasen (I am not sure)" were expected to convey a lower confidence level to users. We investigated the rejection counts, i.e., how many system suggestions were rejected by the participants (maximum: 12 times for each confidence level) in order to check whether each stimulus successfully conveyed the designed confidence level to the participants.

The rejection counts were then analyzed with a 2 × 2 mixed ANOVA (between independent variable: ASE/speech conditions, within independent variable: higher/lower confidences, and dependent variable: rejection counts). The results showed no significant difference in the interaction effect [$F(1, 138) = 0.15$, n.s.] and the main effect of the between independent variable [$F(1, 138) = 0.93$, n.s.], but there was a significant difference in the main effect of the within independent variable [$F(1, 138) = 64.08$, $p < .01$] (Table 1). As a result, the suggestions with a flat ASE and with "dato omoi masu (I bet)" had lower rejection counts, while those with a decreasing ASE and "kamo shiremasen (I am not sure)" had higher counts. We thus confirmed that each stimulus successfully conveyed the assigned confidence levels to the participants appropriately. These results were completely the same with the ones reported in the former studies of ASEs [14, 15].

	ASEs	Speech
Higher Confidence	2.66 (SD = 3.27)	3.27 (SD = 2.91)
Lower Confidence	5.57 (SD = 4.01)	5.91 (SD = 3.89)

Table 1. Average rejection counts for each condition

Results

The response times for each stimulus were analyzed with a 2 × 2 mixed ANOVA (between independent variable: ASE/speech conditions, within independent variable: higher/lower confidences, and dependent variable: response time, Figure 4). The results showed significant differences in the interaction effect [$F(1, 138) = 4.31$, $p < .05$] and in the main effect of the between independent variable [$F(1, 138) = 8.48$, $p < .01$]. The simple main effects of the between and within independent variables were then analyzed, and the results showed significant differences in

the response time for higher and lower confidence stimuli between the ASE and speech conditions [higher confidence: $F(1, 138) = 6.07$, $p < .05$, lower confidence: $F(1, 138) = 9.85$, $p < .01$] and for the speech condition between higher and lower confidence stimuli [$F(1, 138) = 7.75$, $p < .01$], while there were no significant differences in the response time for the ASE condition between higher and lower confidences [$F(1, 138) = 0.02$, n.s.]. To summarize, the participants showed shorter response times for the suggestions with ASEs compared with the suggestions with speech, so our hypothesis that users will show a shorter response time when interpreting ASEs compared with speech sounds used as human-like expressions was verified.

Interestingly, the participants showed different response times depending on the higher and lower confidence information in the speech condition, while they showed no difference in the ASE condition. This suggests that ASEs require a constant cognitive load regardless of whether one interprets higher or lower confidence information, which is a significant advantage of ASEs over speech sounds.

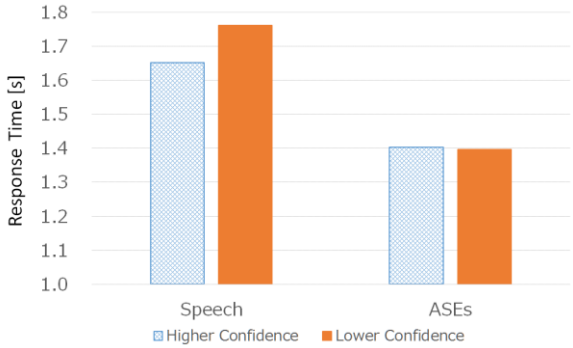

Figure 4. Average response time for each experiment stimulus

DISCUSSIONS AND CONCLUSION

There might be a question of whether or not 24 trials were enough for the participants to get used to the given stimuli (suggestions with speeches). The response times in the speech condition may have gotten closer to those of the ASE condition if the participants had experienced more than 24 trials. Looking separately at the first and latter half of the trials, the average response time gradually decreased in the first half and converged to different times depending on the two conditions, while the standard deviations also gradually decreased in the first half and then converged in the latter half to the same values (Figure 6). This intuition is confirmed by a 2 × 2 mixed ANOVA (between independent variable: ASE/speech conditions, within independent variable: first/latter half, dependent variable: average response time). There are significant differences in both of the main effects [ASE/speech: $F(1, 138) = 7.17$, $p < .01$, first/latter: $F(1, 138) = 46.99$, $p < .01$] (Table 2). From this, we can expect the response time difference of the two conditions to last even after the 24 trials. This strongly supports the idea that interpreting ASEs requires less cognitive load compared with interpreting speech sounds used as human-like expressions.

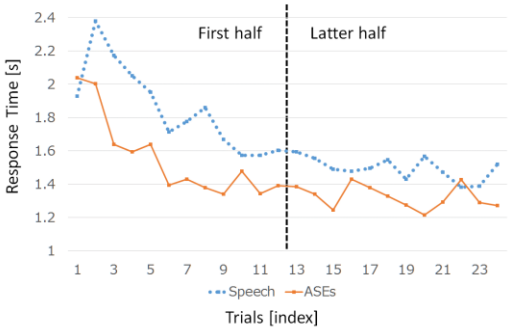

Figure 5. Change in average response time

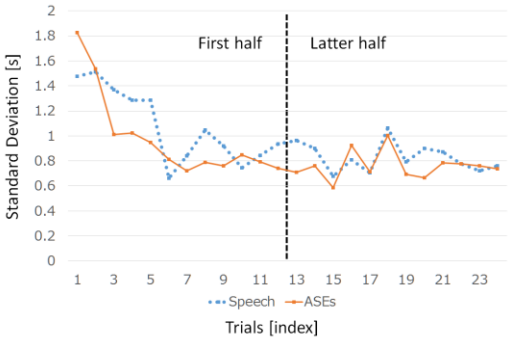

Figure 6. Standard deviations of response time

	ASEs	Speech
First half	1.53 (SD = 0.73)	1.87 (SD = 0.78)
Latter half	1.29 (SD = 0.55)	1.53 (SD = 0.63)

Table 2. Average response times in first and latter halves for each experimental condition

Although some studies reported that "non-speech sounds take longer times to interpret than speech sounds [20, 25]," our study reached the opposite conclusion. We believe that this clearly showed the effectiveness of the ASEs compared with other non-speech sounds like Earcons.

As mentioned in Introduction, previous studies have reported several advantages of ASEs, such as being intuitive for users, being suitable for imperfect systems, and language-independent interpretations being possible. This study successfully added a new advantage, that is, that users showed shorter response times when interpreting ASEs compared with speech sounds used as human-like expressions. Therefore, it should be appropriate to apply ASEs for time-critical applications such as car- or flight-navigation systems and the operation consoles of electric power plants, which need to express emergency information to their operators. For example, in driving situations, less cognitive loads for drivers in interpreting a car's warning can provide more cognitive resources for safe driving. Exploring

how to implement ASEs into such real applications is definitely one of our next targets.

REFERENCES

1. Antifakos, S., Kern, N., Shiele, B., and Schwaninger, A. Towards Improving Trust in Context Aware Systems by Displaying System Confidence, In *Proc. MobileHCI'05*, ACM Press (2005), 9-14.
2. Beatty, J. Task-evoked pupillary responses, processing load, and the structure of processing resources, *Psychological Bulletin 91*, (1982), 276-292.
3. Bellotti, V. and Edwards, K. Intelligibility and accountability: Human considerations in context-aware systems, *Human-Computer Interaction 16*, 2 (2001), 193-212.
4. 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.
5. Blattner, M. M., Sumikawa, D. A., and Greenberg, R. M. Earcons and Icons: Their Structure and Common Design Principles, *SIGCHI Bull. 21*, 1 (1989), 123-124.
6. Cai, H. and Lin, Y. Tuning Trust Using Cognitive Cues for Better Human-Machine Collaboration, In *Proc. HFES2010*, pp. 2437-2441(5).
7. Feng, J. and Sears, A. Using Confidence Scores to Improve Hands-Free Speech Based Navigation in Continuous Dictation Systems, *ACM Transactions on Computer-Human Interaction 11*, 4, ACM Press (2004), 329-356.
8. Fredericks, T. K., Choi, S. D., Hart, J., Butt, S. E., and Mital, A. An investigation of myocardial aerobic capacity as a measure of both physical and cognitive workloads, *International Journal of Industrial Ergonomics 35 (12)*, (2005), 1097-1107.
9. Horvitz, E., and Barry, M. Display of information for time-critical decision making, In *Proc. 11th Conf. on Uncertainty in Artificial Intelligence*, Morgan Kaufmann (1995), 296-305.
10. Horvitz, E. Principles of mixed-initiative user interfaces, In *Proc. CHI'99*, ACM Press (1999), 159-166.
11. Hussain, S., Chen, S., Calvo, R. A., and Chen, F. Classification of cognitive load from task performance & multichannel physiology during affective changes, In *Proc. MMCogEmS: Inferring Cognitive and Emotional States from Multimodal Measures (ICMI 2011 Workshop)* (2011).
12. Kahneman, D. Attention and effort, Prentice-Hall, New Jersey, (1973).
13. Katehakis, M, N. and Veinott, A, F, Jr. The Multi-Armed Bandit Problem: Decomposition and Computation, *Mathematics of Operation Research 12*, 2 (1987), 262-268.
14. Komatsu, T., Yamada, S., Kobayashi, K., Funakoshi, K., and Nakano, M. Artificial Subtle Expressions: Intuitive Notification Methodology for Artifacts, In *Proc. CHI'10*, ACM Press (2010), 1941-1944.
15. Komatsu, T., Kobayashi, K., Yamada, S., Funakoshi, K., and Nakano, M. How Can We Live with Overconfident or Unconfident Systems?: A Comparison of Artificial Subtle Expressions with Human-like Expression, In *Proc. CogSci2012* (2012), 1816-1821.
16. Komatsu, T., Kobayashi, K., Yamada, S., Funakoshi, K., and Nakano, M. Investigating Ways of Interpretations of Artificial Subtle Expressions Among Different Languages: A Case of Comparison Among Japanese, German, Portuguese and Mandarin Chinese, In *Proc. CogSci2015* (2015), 1159-1164.
17. Komarov, S., Reinecke, K., and Gajos, K. Z. Crowdsourcing performance evaluations of user interfaces, In *Proc. CHI'13*, ACM Press (2013), 207-216.
18. Liu, B. and Lane, I. Joint Online Spoken Language Understanding and Language Modeling with Recurrent Neural Networks, In Proc. *SIGDIAL'16*, (2016), 22-30.
19. Navon, D. and Gopher, D. On the economy of the human-processing system, *Psychological Review 86*, (1979), 214-255.
20. Noyes, J. M., Hellier, E., and Edworthy, J. Speech warnings: a review, *Theoretical Issues in Ergonomics Science 7*, 6 (2006), 551-571.
21. Mesnil, G., Dauphin, Y., Yao, K., Bengio, Y., Deng, L., Hakkani-Tur, D., He, X., Heck, L., Tur, G., Yu, D., and Zweig, G. Using Recurrent Neural Networks for Slot Filling in Spoken Language Understanding, IEEE Transactions on Audio, Speech, and Language Processing 23 (3), (2015), 530-539.
22. Ogawa, A. and Nakamura, A. Joint estimation of confidence and error causes in speech recognition, *Speech Communication 54*, 9 (2012), 1014-1028.
23. Paas, F. G. W. C. and Van Merrienboer, J. J. G. The Efficiency of Instructional Conditions: An Approach to Combine Mental Effort and Performance Measures, *Human Factors: the Journal of the Human Factors and Ergonomics Society 35 (4)*, (1993), 737-743.
24. Sweller, J. Cognitive Load During Problem Solving: Effects on Learning, *Cognitive Science 12*, (1998), 257 - 285.
25. Vilimek, R., and Hempel, T. Effects of speech and non-speech sounds on short-term memory and possible implications for in-vehicle use, In *Proc. ICAD 2005*, (2005).