Investigating Ways of Interpretations of Artificial Subtle Expressions Among Different Languages: A Case of Comparison Among Japanese, German, Portuguese and Mandarin Chinese

Takanori Komatsu (tkomat@meiji.ac.jp)
Department of Frontier Media Science, Meiji University
4-21-1 Nakano, Tokyo 1648525 Japan

Rui Prada (rui.prada@tecnico.ulisboa.pt)
INESC-ID, Av. Prof. Aníbal Cavaco Silva,
Porto Salvo 2744016, Portugal

Kazuki Kobayashi (kby@shinshu-u.ac.jp)
Department of Computer Science and Engineering, Shinshu University
4-7-1 Wakasato, Nagano 3808553, Japan

Seiji Yamada (seiji@nii.ac.jp)
National Institute of Informatics/SOKENDAI/Tokyo Institute of Technology
2-1-2 Hitotsubashi, Tokyo 1018430 Japan

Kotaro Funakoshi (funakoshi@jp.honda-ri.com)
Honda Research Institute Japan, Co., Ltd.
8-1 Honcho, Wako, Saitama 3500188, Japan

Mikio Nakano (nakano@jp.honda-ri.com)
Honda Research Institute Japan, Co., Ltd.
8-1 Honcho, Wako, Saitama 3500188, Japan

Abstract

Up until now, several studies have shown that a speech interface system giving verbal suggestions with beeping sounds that decrease in pitch conveyed a low system confidence level to users intuitively, and these beeping sounds were named "artificial subtle expressions" (ASEs). However, all participants in these studies were only Japanese, so if the participants' mother tongue has different sensitivity to variations in pitch compared with Japanese, the interpretations of the ASEs might be different. We then investigated whether the ASEs are interpreted in the same way as with Japanese regardless of the users' mother tongues; specifically we focused on three language categories in traditional phonological typology. We conducted a web-based experiment to investigate whether the ways speakers of German, Portuguese (stress accent language), Mandarin Chinese (tone language) and Japanese (pitch accent language) interpret the ASEs are different or not. The results of this experiment showed that the ways of interpreting did not differ, so this suggests that these ways are language-independent.

Keywords: Artificial subtle expressions (ASEs); tone language; pitch accent language; stress accent language

Introduction

Although there is little hope that speech interface systems will ever be perfectly reliable (Bellotti & Edwards, 2001; Ogawa & Nakamura, 2012), people's interaction with such non-perfect systems has only been analyzed sparsely (Higashinaka et al, 2006). Recently, 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 (Benzeghibaa et al., 2007; Feng & Sears, 2004; Horvitz, 1999; Parasuraman, 1997). For example, Antifakos et al. (2005) showed that users adapt to a system easily if the system's confidence is displayed on a computer screen. Horvitz & Barry (1995) proposed a context-aware system that can estimate the expected value of revealed information to enhance computer displays for time-critical applications. Cai & Lin (2010) experimentally showed how expressing the level of confidence for such system to indicate whether the system's represented information is accurate or not to users plays an important role in improving both the users' performance and their subjective impressions. Therefore, expressing a system's confidence to users is an important requirement for user interfaces. Most of these studies used actual human-like expressions to express their confidence level to users, e.g., speech sounds from speakers or character information on computer displays.

In contrast with the above approaches, Komatsu et al. (2010a, 2010b) proposed using artificial subtle expressions (ASEs) as machine-like expressions used to convey a

system's confidence level to users. Specifically, they proposed two simple beeping sounds used as ASEs: a flat sound (flat ASE) and a sound with a decreasing pitch (decreasing ASE). These ASEs were added 0.2 seconds after the system's verbal suggestions. They then showed that suggestions made with decreasing ASEs conveyed a low system confidence level to users (Figure 1).

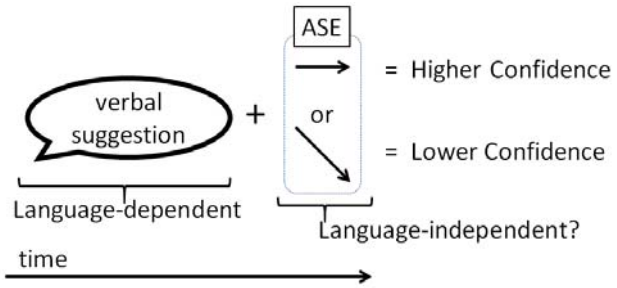

Figure 1: Artificial Subtle Expressions (ASEs).

These experiments were conducted only in Japan, and all of the participants were Japanese, so there is some possibility that the sensitivity of the participants in their mother tongue to variations in pitch might affect the ways ASEs are interpreted; for example, if the participants have different sensitivity to variations in pitch compared with Japanese, the interpretations of decreasing ASEs might be different. In terms of the sensitivity to such pitch variation, the three language categories shown in Table 1 are proposed in traditional phonological typology (Hirst & Di Cristo, 1998; McCawley, 1978; O'Grady et al., 1997; Van der Hulst & Smith, 1998). "Pitch accent language" means that variations in pitch can be used to differentiate words, "tone language" means that different tones change the meaning of the words (more sensitive to variations in pitch compared with pitch accent language), and "stress accent language" means that only the variations in power are used for expressing an accent (less sensitive compared with pitch accent languages).

Table 1: Language categorization in phonological typology.

Type	Example of languages
Tone language	Mandarin Chinese, Thai, Vietnamese
Pitch accent language	Japanese, Swedish
Stress accent language	English, German, Spanish, Portuguese

Therefore, to investigate whether the ASEs are interpreted in the same way as with Japanese (pitch accent language) regardless of the users' mother tongues, it is necessary to investigate how the ASEs are interpreted by participants whose mother tongues are categorized as the remaining two language types, e.g., tone language and stress accent language. Specifically, we focused on Mandarin Chinese as the tone language and German and Portuguese as the stress accent language (Table 1). The reason that we focused on two stress accent languages is the different positions of stress in a word. In general, German has a strong stress on the first syllable, while Portuguese does on the second or third last syllable (Hirst & Di Cristo, 1998).

We thus conducted a web-based experiment to investigate whether the ways speakers of German, Portuguese, Mandarin Chinese, and Japanese interpret ASEs are different or not. This investigation would clarify whether these ways are language-independent or not and will determine whether the ASEs can be applied to various kinds of speech interface systems regardless of the users' spoken language or mother tongues, while most current techniques for speech interface systems are obviously language-dependent.

Experiment

Settings

We conducted a web-based experiment to investigate the effects of the participants' mother tongue on their interpretation of the ASEs. 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).

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 top image of Figure 2) told them which mound it expected the coin to be in by using speech with the ASEs. 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 middle image of Figure 2). The participants were informed of their total numbers of coins only after they finished all 24 trials.

Stimuli

The navigation system used speech with the ASEs to tell participants the expected position of the coin. We prepared three different pieces of speech and two different ASEs. This means that the system could present six different stimuli for the participants. The navigation system used English speech sounds to suggest to the participants the expected location of the coin, that is, "number one,"

"number two," and "number three." These speech sounds were generated by AT&T Natural Voices[1].

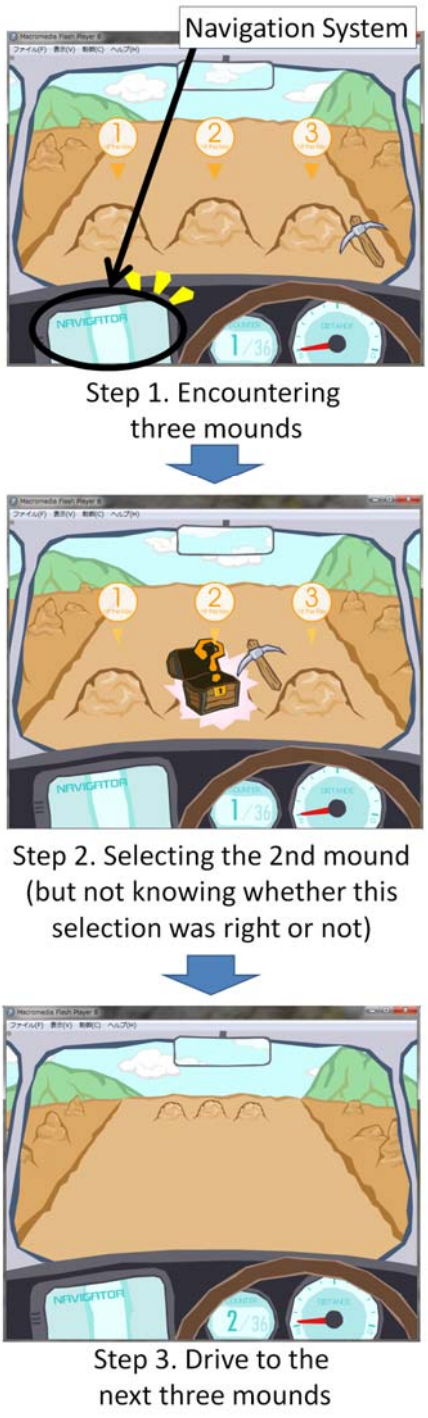

Figure 2. Driving treasure hunting video game

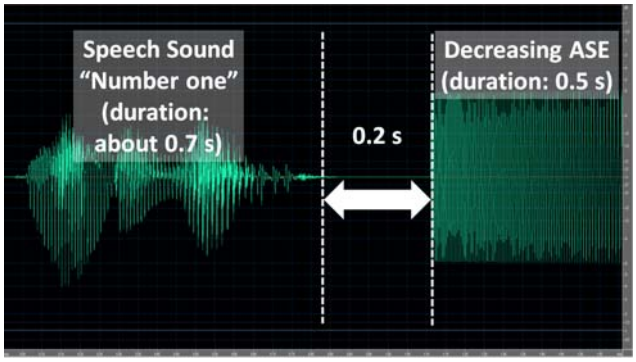

Figure 3. Speech waveform of verbal suggestion "number one" with decreasing ASE

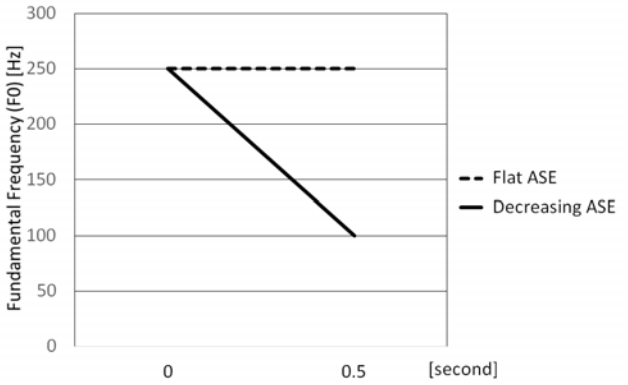

Figure 4. Flat and decreasing ASEs

0.2 seconds after the end of the speech sounds, one of the two ASEs were played (Figure 3). These two ASEs were triangle wave sounds 0.5 seconds in duration, but their inflection patterns of pitch were different; that is, one was a flat ASE (onset fundamental frequency (F0): 250 Hz and end F0: 250 Hz), and the other was a decreasing ASE (onset F0: 250 Hz and end F0: 100 Hz) (Figure 4). The former studies (Komatsu et al., 2010a, 2010b) already showed that suggestions made with decreasing ASEs conveyed a low system confidence level to users. The interval between the suggestions and the ASEs (0.2 seconds) and the duration of the ASEs (0.5 seconds) was the same as in the former studies.

Participants

117 volunteers (73 men and 44 women; 20 - 45 years old, mean age of 24.91) participated. These participants voluntary responded to a call for participants from the authors. Out of the 117 participants, 44 participants' mother tongue was Portuguese (Portuguese group, nationality: Portugal), 26 was Japanese (Japanese group, nationality: Japanese), 23 was German (German group, nationality: Germany, Austria, and Switzerland), and 24 was Mandarin Chinese (Chinese group, nationality: China).

All the instructions in the experimental system were written in English so that all the participants could

[1] http://www2.research.att.com/~ttsweb/tts/demo.php

understand these English descriptions. Therefore, there was some possibility that the participants would react to the given suggestions as if their mother tongue were English, but the verbal suggestions used are quite simple English phrases like "number one", and Komatsu et al. (2010b) reported that participants interpret the meanings of ASEs intuitively and unconsciously, so we assumed that their interpretation of ASEs was not significantly affected by their English skills but purely by their mother tongue.

Procedure

We used a web-based experiment system for participants to play the treasure hunting video game and to record the participants' behavior in regard to selecting which mound contained the coin depending on the given suggestions. First, the system displayed a consent form and instructions for the experiment. These instructions never mentioned or explained the ASEs to the participants. 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 driving treasure hunting video game. Among the 24 trials, the system expressed all 6 stimuli 4 times (the flat ASEs 12 times and the decreasing ASEs 12 times, see Figure 5). The order in which the stimuli were expressed was randomized.

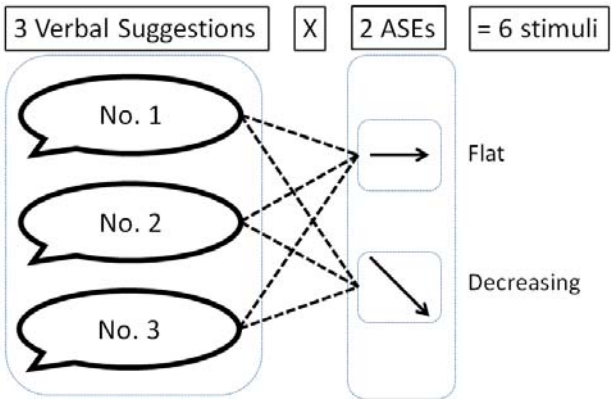

Figure 5. Six stimuli consist of three different pieces of speech and two different ASEs

Results

To investigate the effects of the different ASEs on the participants' behavior in terms of how often they accepted or rejected the system's suggestions in each group, we then counted the rejection count, indicating how many of system's suggestions the participants actually rejected. The average rejection counts of the 12 flat ASEs and 12 decreasing ASEs for the four participant groups are summed in Table 2.

Table 2: Rejection counts of flat and decreasing ASEs for four language groups.

Language	12 flat ASEs	12 decreasing ASEs
Portuguese (n = 44)	3.27 (SD = 3.03)	6.45 (3.73)
Japanese (n = 26)	3.62 (3.22)	6.27 (4.01)
German (n = 23)	3.69 (3.17)	5.83 (3.78)
Mandarin Chinese (n = 24)	3.67 (2.93)	6.13 (2.88)

These rejection counts were analyzed by using a 2 × 4 mixed plan ANOVA (within-participant independent variable: types of ASEs, flat/decreasing, between-participant independent variable: language group, Portuguese/German/ Japanese/Chinese, dependent variable: rejection counts). The results of the ANOVA showed that there was no significant differences in the interaction effect [$F(3,114) = 0.44$, n.s.] and in the main effect of the between-participant independent variable (four language groups) [$F(3,114) = 0.02$, n.s.], but there was a significant difference between the within-participant independent variables (two ASEs) [$F(1,114) = 44.02$, $p < .01$, effect size: $\eta^2 = 0.13$] (Figure 6).

To sum up, the system's suggestions with the decreasing ASEs showed significantly higher rejection counts compared with those with the flat ASEs, regardless of the participants' mother tongue. Therefore, the ways speakers of Portuguese, Japanese, German, and Mandarin Chinese interpreted the ASEs did not differ from each other, so this result suggests that the ways are language-independent.

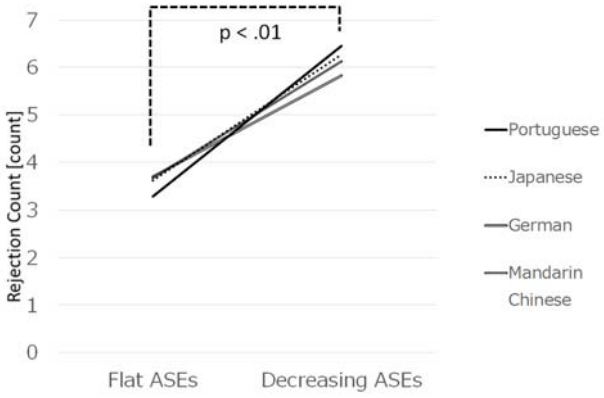

Figure 6. Rejection counts for flat and decreasing ASEs and observed significant differences among these factors

Discussion

Japanese, German, Portuguese, Mandarin Chinese, and what else?

In this paper, we succeeded in clearly showing that the ways of interpreting ASEs did not differ among the participants whose mother tongues are a tone language (Mandarin Chinese), pitch accent language (Japanese), and stress accent language (Portuguese and German). This result suggests that the ways are language-independent. However, we have to tackle several consecutive experiments in order to conclude that the ways are "perfectly" language-independent. One is conducting experiments by recruiting participants whose mother tongues are English, Spanish, Hindi, and Arabic because these populations total nearly one billion, and the other is conducting an experiment with those who cannot understand English at all because our experimental system was implemented in English and the participants in this study can understand English as a foreign language. Therefore, we could not perfectly exclude the effects of English in the results of this paper. To tackle with this issue, we are planning to implement an experimental system with showing the participants' mother tongue. We believe the results of these consecutive studies would compensate for the effectiveness of the ASEs and will lead to the above strong conclusions.

Limitations and future direction of studies about ASEs

In this experiment, we utilized the ASEs in a gaming environment in which participants simply needed to accept or reject the system's suggestions. Currently, the ASEs is designed to convey only high or low confidence to the users, so the application range of the ASEs seems to be quite limited. Although conveying high/low confidence to users is an abstract gaming task, it is quite important and effective for systems that need to tell users what they should do next, such as car navigation systems giving route guidance like "turn left" or "enter highway #I-8."

Currently, we are wondering whether this simple expression ASEs is also effective for systems that are required to give much more complex information to users, such as those that need to express a degree of confidence level not simply expressed with "higher" or "lower," like information retrieval systems (Sanderson & Zobel, 2005) or recommendation systems (Re Roure & Shadbolt, 2001). In such more complex systems, we speculate that not only decreasing ASEs but other inflection patterns of ASEs or ranges of pitch variation of ASEs will have specific meanings. We are now planning to use other kinds of gaming environments to handle much more complex and flexible interaction with users. The results of such experiments in these more complex systems should expand the application range of ASEs.

ASEs are proposed as simple and intuitive expressions for users to convey the levels of the confidence of the artifacts. Although several studies already clarified the various effectiveness of ASEs, e.g., preciseness and robustness (Komatsu et al, 2010a, 2010b), we have not yet investigated how much the users' cognitive loads are utilized for interpreting the ASEs. We expect that the interpretation of the ASEs would consume the less cognitive loads compared to the interpretations of the verbal expressions that convey the confidence level of the artifacts (e.g., "I am 80% confident" or "you MUST follow my suggestion"). We also have to tackle this unsolved issue.

Conclusions

This paper succeeded in clearly showing that the ways of interpreting ASEs did not differ among the participants whose mother tongues are a tone language (Mandarin Chinese), pitch accent language (Japanese), and stress accent language (Portuguese and German). This result suggests that the ways are language-independent, while most techniques for speech interface systems are obviously language-dependent. We believe that these ASEs can then be applied to various kinds of speech interface systems, e.g., car navigation systems or speech recognition systems in mobile devices, regardless of the users' mother tongues or spoken languages.

Acknowledgments

This work was supported by KAKENHI (25330319) as Grant-in-Aid for Scientific Research (C) by The Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan, by Fundação para a Ciência e a Tecnologia under project PEst-OE/EEI/LA0021/2013, and by joint research with Honda Research Institute Japan and with National Institute of Informatics, Japan.

References

Antifakos, S., Kern, N., Shiele, B. & Schwaninger, A. (2005). Towards Improving Trust in Context Aware Systems by Displaying System Confidence, *Proceedings of the 7th International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI'05)*, pp. 9-14.

Bellotti, V. & Edwards, K. (2001). Intelligibility and accountability: Human considerations in context-aware systems, *Human-Computer Interaction, 16 (2)*, 193-212.

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. & Wellekensa, C. (2007). Automatic speech recognition and speech variability: A review, *Speech Communication, 49 (10-11)*, 763-786.

Cai, H. & Lin, Y. (2010). Tuning Trust Using Cognitive Cues for Better Human-Machine Collaboration, *Proceedings of the 54th Annual Meeting of the. Human Factors and Ergonomics Society (HFES2010)*, pp. 2437-2441(5).

Feng, J. & Sears, A. (2004). Using Confidence Scores to Improve Hands-Free Speech Based Navigation in Continuous Dictation Systems, *ACM Transactions on Computer-Human Interaction, 11 (4)*, 329-356.

Higashinaka, R., Sudoh, L. & Nakano, M. (2006). Incorporating Discourse Features into Confidence Scoring of Intention Recognition Results in Spoken Dialogue Systems, *Speech Communication 48 (3-4)*, 417-436.

Hirst, D. & Di Cristo, A. (Eds.) (1998). *Intonation systems: A survey of twenty languages*, Cambridge, UK: Cambridge University Press.

Horvitz, E. & Barry, M. (1995). Display of information for time-critical decision making, *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, pp. 296-305.

Horvitz, E. (1999). Principles of mixed-initiative user interfaces, *Proceedings of the 17th ACM Conference on Human Factors in Computing Systems (CHI'99)*, pp. 159-166.

Komatsu, T., Yamada, S., Kobayashi, K., Funakoshi, K. & Nakano, M. (2010a). Artificial Subtle Expressions: Intuitive Notification Methodology for Artifacts, *Proceedings of the 17th ACM Conference on Human Factors in Computing Systems (CHI2010)*, pp. 1941-1944.

Komatsu, T., Yamada, S., Kobayashi, K., Funakoshi, K. & Nakano, M. (2010b). Proposing Artificial Subtle Expressions as an Intuitive Notification Methodology for Artificial Agents' Internal States, *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society (CogSci2010)*, pp. 447-452.

McCawley, J. D. (1978). What is a tone language, In V, Fromkin (Eds.), *Tone: a linguistic survey*, 113-131, Waltham, MA: Academic Press.

Ogawa, A. & Nakamura, A. (2012). Joint estimation of confidence and error causes in speech recognition, *Speech Communication, 54 (9)*, 1014-1028.

O'Grady, W., Dobrovolsky, M. & Aronoff, M. (1997) *Contemporary Linguistics (Third Edition)*, New York: St. Martin's Press.

Parasuraman, R. (1997). Human use and abuse of Automation, In M. Mouloua and J. Koonce (Eds.), *Human-Automation Interaction: Research and Practice*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Re Roure, D. C. & Shadbolt, N. R. (2001). Capturing knowledge of user preferences: Ontologies in recommender systems, *Proceedings of the 1st Conference on Knowledge Capture*, pp. 100-107.

Sanderson, M. & Zobel, J. (2005). Information retrieval system evaluation: effort, sensitivity, and reliability, *Proceedings of the 28th ACM SIGIR International Symposium on Information Retrieval,* pp. 162-169.

Van der Hulst, H. & Smith, N (Eds.) (1998). *Autosegmental studies in pitch accent*, Hawthorne, NY: Foris publication.