Dynamic estimation of emphasizing points for user satisfaction evaluations

Yoshimasa Ohmoto(ohmoto@i.kyoto-u.ac.jp)
Department of Intelligence Science and Technology
Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501 Japan

Takashi Miyake (miyake@ii.ist.i.kyoto-u.ac.jp)
Department of Intelligence Science and Technology
Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501 Japan

Toyoaki Nishida (nishida@i.kyoto-u.ac.jp)
Department of Intelligence Science and Technology
Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501 Japan

Abstract

When many factors must be considered for decision-making, people dynamically change their emphasizing points, along with their understanding of these factors and the relationships between them. In previous work, we proposed a method to dynamically estimate emphasizing points (DEEP) based on utterances, physiological indices, and proposal selections. To evaluate this method in actual interactions, we conducted controlled WoZ (Wizard of Oz) experiments using Embodied Conversational Agents (ECAs), which interactively provide controlled information for decision-making. Using ECAs, we compare our method to an existing method, which estimates emphasizing factors through the " gradual method ". We confirm that our method can accurately estimate dynamic changes of emphasizing points, and that participants were more satisfied with the final proposal from the ECA that used DEEP.

Keywords: verbal and nonverbal behavior; physiological indices; preferential structure estimation.

Introduction

When many factors must be considered for decision-making, we dynamically and interactively change the factors that we emphasize (which we call "emphasizing points"). We also change our understanding of these factors and relationships between them. For example, in travel planning, we have to synthetically consider factors, such as place, budget, members, and schedule. We often make such plans interactively with our friends and travel agency staff.

The interaction between conversational partners influences how we understand the factors and the relationships between them during the decision-making process. Therefore, their emphasizing points are often dynamically changed when faced with new information. However, the important factors may not only be the most recent points emphasized, but the process of interaction may also change the emphasizing points. People have to re-estimate the changes in their emphasizing points throughout the interaction.

In interactive decision-making with dynamic changes to emphasizing points, humans provide active demands and passive responses through verbal expressions, nonverbal reactions, proposal selection, and physiological state (Ohmoto, Kataoka, Miyake, & Nishida, 2011). In the previous work, we analyzed the interaction process, and verbal and nonverbal behavior during the interaction to propose an estimation method of interaction using utterances, nonverbal behavior, physiological indices, and proposal selections.

The purpose of this study is to evaluate whether our proposed method based on dynamic estimation of emphasizing points (DEEP) is useful to participants in interactive decision-making and whether the proposed method can provide satisfactory proposals for participants. To test this method, we used Embodied Conversational Agents (ECAs), because it is difficult for human agents to achieve rigorously controlled interaction with participants based on our proposed method. Specifically, we conducted an experiment that compares the results of interactive decision-making with two types of ECAs; one provided proposals based on our method and another based on an existing method that gradually estimates emphasizing points based on verbal expressions and proposal selections.

The paper is organized as follows: Section 2 discusses work on interactive systems. Section 3 briefly explains DEEP, which dynamically estimates emphasizing points. Section 4 describes the experiment for comparing two types of estimation methods and then presents the results. Section 5 discusses the achievements and limitations of the proposed method. Section 6 concludes and discusses future work.

Related work

Some researchers have developed systems that can provide proposals to satisfy user's demands. These systems gradually estimate user's demands throughout the interaction.

Kitamura et al. (Kitamura et al., 2008) developed the "Laddering" Search Service System that matches users queries with search targets by communicating with users throughout the interview. They assume that user's emphasizing points do not change during the interaction.

Aydogan et al. (Aydogan & Yolum, 2007) proposed an architecture in which both consumers and producers use a

shared ontology to negotiate services. Through repetitive interactions, the provider accurately learns consumers' needs to provide better-targeted offers. The system learns consumers' needs over long-term interactions.

Kurata (Kurata, 2010) proposed a computer-aided tour planning system. The system provides several tour plans and asks the user to provide feedback. The feedback is utilized by the system for inferring the user's preferences and then for revising tour plans. This cycle is repeated until the user is satisfied with the final plan, with the hopes that this method gradually leads to a more satisfying experience of computer-aided tour planning. The system can then estimate user's emphasizing points. However, the user has to manually change emphasizing points when the user wants to change her/his emphasizing points during the interaction. Moreover, the user cannot modify their emphasizing points when he/she does not have knowledge about the planning.

Previous work revealed that user demands and needs could gradually be estimated through repetitive interactions. However, most of the research did not consider that user's demands and needs could change throughout the interaction. In contrast, we assume that emphasizing points can change over the interaction and we dynamically estimate these changes. We focus not only on active demands verbally expressed and proposal selections, but also on passive responses expressed by backchanneling, and nonverbal reactions.

It is, however, difficult to estimate human internal states through nonverbal information, especially when passively interacting with others. Therefore, we use physiological indices for estimating human internal states during interaction. There are various studies on estimating human internal states by measuring physiological indices (e.g. (Iwaki, Arakawa, & Kiryu, 2008)). There are also several studies that use these measured physiological indices for effective human-agent interaction.

Bosma et al. (Bosma & Andre, 2004) proposed a method that takes into account users' emotional state to disambiguate dialogue acts. They restrict to pedagogical agents that offer a text-based natural language interface for assisting the user in text communication. They estimated levels of arousal and valence by using physiological indices: skin conductivity response (SCR), heart rate, muscle activity, and respiration rate.

Prendinger and Ishizuka (Prendinger & Ishizuka, 2005) developed an interview agent which takes physiological data (skin conductance and electromyography) of users in real-time, interprets the data into emotions, and addresses the user's affective states in the form of emphatic feedback. In addition, they evaluated the agent by using SCR and heart rate. The empathic feedback has a positive effect on the interviewee's stress level while hearing the question.

As mentioned above, physiological indices are useful for estimating human internal states in interaction even when users passively interact with others. The proposed method uses physiological indices, SCR, electrocardiograms (LF/HF values), and skin temperature of fingers, to detect mental stress, such as pleasure, excitement, and tension. The method estimates emphasizing points by using these physiological indices, as well as verbal expressions, and nonverbal responses.

We have discussed the achievements and limitations of previous work related to our objective of estimating emphasizing points for interactive decision-making. Because of the difficulty in detecting passive responses during interactions, most prior work estimated user demands and needs gradually through repetitive interactions that required active demands from users. Therefore, we propose a method that dynamically estimates emphasizing points by using physiological responses, which could detect human internal states even during a passive interaction, *in addition to* verbal expressions, and nonverbal responses. In this study, we apply the proposed method to actual interactions and experimentally evaluate whether proposals that use physiological responses are useful for participants' decision-making and for achieving satisfactory results in the interaction.

Dynamically estimating emphasizing points

For our purpose, we conducted preliminary analyses to elicit useful information for dynamically estimating emphasizing points (DEEP) in human-human interaction (Ohmoto et al. 2011). As a result of the analyses of videos and physiological indices, we could suggest a method to DEEP which is explained next subsection. We proposed a method to DEEP based on the observation of human-human interaction in preliminary analyses, so, we think that the proposed method is one of methods realizing DEEP. In this section, we briefly explain the proposed method to DEEP based on verbal reactions, body movements, and physiological indices, when participants are given two proposals and asked for his/her selection and demands.

DEEP, in this paper, is applied to the situation in which many factors, including unknowns, for must be considered for decision-making. In this situation, a user interacts with a system based on DEEP and the system advises some useful proposals for user's decision-making. A proposition process in an interaction is as follows: First, the two most appropriate proposals at that point are explained from a DEEP system. After the proposition, the system asks the user what his/her demands were and which proposal is better. The DEEP system pays attention to the user's reactions and answers during the explanation and questions. The system then estimates the emphasizing points. The user repeats this process until one of the propositions satisfies the user's end goal.

Overview of DEEP

The degree of emphasis for an emphasizing point is rated on a scale from zero to five. The rating is changed based on the following three factors during the explanation.

- **Verbal reactions**
  Either of the two following reactions occurs.
  - Listed words appear in answers or demands.

– The participant provides backchanneling phrases, which express acknowledgement, surprise, or understanding, such as ”ah,” ”oh,” ”aha,” ” I see,” and ”I understand.”

- **Body movements**
  The participant repeatedly nods three times or more.

- **Physiological indices**
  Either of the two following responses occur (refer to (Miyata, 1998), (Lin, Omata, Hu, & Imamiya, 2005), (Iwaki et al., 2008) and (Nakazono, Hada, Ataka, Tanaka, & Nagashima, 2008)).

  – SCR increases more than 10% compared to resting level.
  – LF/HF value (electrocardiograph measurement) is more than 6.0.

Verbal reactions, body movements, and physiological indices, are used as criteria for determining when a new factor is discovered and should be emphasized, and for determining when a user's degree of emphasis of a particular factor increases or decreases.

Rules for changing estimated emphasizing points during explanation The estimated emphasizing points are changed by the participant's responses when a DEEP system explains the proposals.

- **Discovery of a new factor to be emphasized**
  When any one of the three criteria appears during an explanation, the system decides that the factor should be slightly emphasized, and increases the degree of emphasis from zero to two. When any two or three criteria are present, the system increases the emphasis from zero to three.

- **Increasing or decreasing degree of emphasis**
  When any one of the three criteria appears, the system decides that the factor should be emphasized, and increases the emphasis of the factor by one. When there are physiological reactions, but no verbal reactions, or body movements, the system decides that the factor should be emphasized less, and decreases the emphasis of the factor by one.

Rules for changing estimated emphasizing points from active demands The system asks whether a user has any demands. From the user's response, the system determines what the user ' s demands are and what changes there are to emphasizing points. The system uses assumed keywords in the user ' s response to determine demands and changes to demands. Assumed keywords are words that express assumed emphasizing points, demands, and basic words necessary to capture demands. Words that are not expected to be included in answers are ignored.

- **Discovery of new factors to be emphasized**
  When the emphasis degree of the discovered factor is zero, the system increases the degree of emphasis from zero to three.

- **Increasing or decreasing degree of emphasis**
  When the emphasis of the discovered factor is greater than zero and the system decides that the factor should be increased, the system increases the degree by one. When the system decides that the emphasis of the factor should be decreased and the degree is greater than zero, the system decreases the degree by one.

Deciding a better proposal by the user's choice between the two proposals Given two proposals, the system asks the user which is better. If the proposal satisfied the user's end goal, that is the final proposal. If not, based on the answer, the system determines which proposal more satisfies the user or decides either that both proposals equally satisfy or that neither proposal is satisfactory. When the system determines that both proposals equally satisfy the user, the proposal in which the lowest skin temperature was recorded is regarded as better. When the system determines that neither proposal satisfieds the user, the system does nothing.

Selecting the next step based on DEEP results

According to the criteria mentioned above, changes to user ' s emphasizing points are estimated after the proposals are given and data is collected from the user ' s reactions and response. After the estimation, the next two proposals are selected based on the estimation results.

The next proposals are selected using a table of orthogonal arrays in advance. Orthogonal arrays are a special set of Latin squares, which can be used to estimate main effects using only a few experimental runs. From the table, the two proposals that most satisfy user's emphasizing points are picked. When many proposals in the table can satisfy a user ' s emphasizing points, the two proposals nearest to the best proposal for a user's choice are selected. When neither proposal will satisfy the emphasizing points, the two proposals furthest from the previous proposition are selected. The distances of proposals are calculated by cosine similarity.

Experiment

The purpose of this experiment was to investigate whether the DEEP method could accurately estimate emphasizing points in which many factors, including unknown factors, must be considered for decision-making. In the experiment, we used human-like virtual agents (ECAs) to strictly control the verbal and nonverbal expressions of the agent, which could affect user ' s impressions of the proposals presented. The ECAs were operated by a WoZ (Wizard of Oz) interface because accurate voice recognition can be difficult. The proposed method was compared with the gradual method, which was discussed above, and is described in more detail below.

Task

Participants were asked to design a mobile robot using a robot parts catalogue. Each participant interacted with an experimenter for two sessions, in which they designed a different robot that achieved different tasks. The participant could

change the design concept of the robot during the session without informing the change to the experimenter. The task had 23 criteria that the robot must meet and there were various ways to design robots that realize the same purpose. Examples purposes in Situation A were ”taking photos of beautiful scenery” and ”introducing old temples and shrines,” while in Situation B, examples purposes were ” a mountain climbing race” and ”a city obstacle race.”

The gradual method compared with DEEP

We compared the DEEP method with gradual method. In the gradual method, the ECA provides the two proposals nearest to the best proposal of the user. When the user decides that neither proposal will suffice, the two proposals furthest from the last two proposals. This method only uses user's selection between the two proposals and gradually approaches a satisfactory proposal. The method does not pay attention to the dynamic changes of user's emphasizing points during the interaction. Therefore, only the user's actual choice is taken into account. This method can provide a better proposal than previous one in most cases. This is a better point than the DEEP method. This method was regarded as a modified version of work by Kurata (Kurata, 2010).

Outline of WoZ

The experimenter entered into the system data that contained verbal reactions, body movements, and physiological indices, because we could not robustly capture this data in real-time. Each ECA generated verbal and nonverbal behavior that had been previously designed by the experimenter based on the expected reactions.

Both ECAs accepted the results of user's choice. In addition, the ECA with DEEP accepts data as was described in previous section. Verbal reactions and body movements are determined via visual observation. Physiological indices were automatically measured and the experimenter annotated which words or explanations may have triggered the physiological responses. Each ECA used the entered data to decide the proposals presented in the next proposition.

Experimental settings

The experimental setting is shown in Figure 1. The participant sat in front of a 100-inch screen displaying the ECA. The experimenter sat out of view of the participant and entered the stimuli via a WoZ interface. Two video cameras recorded the participant's behavior; one was placed on the screen for recording the participant's behavior, and another was placed behind the participants for recording the screen. The participant's voice was recorded by microphones. Polymate was used to measure SCR, the electrocardiogram, and skin temperature of fingers. The experimenter instructed the participant to keep their left arm on an armrest.

Participants

26 students (20 males and 6 females) participated in the experiment. They were undergraduate students from 18 to 25 years old (an average of 20.6 years old). They did not know about robots but they were in science course. All of them interacted with both of ECA with DEEP and without DEEP.

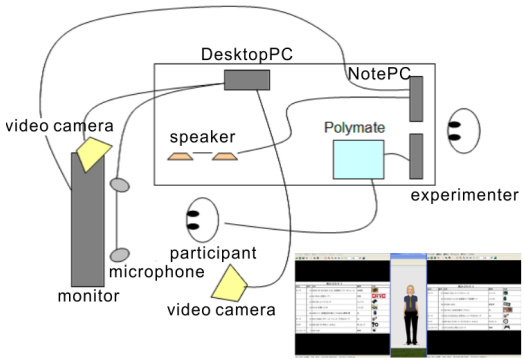

Figure 1: Experimental settings

Table 1: T-test results for accuracy in estimating emphasizing points

	proposed	gradual
average	2.1	1.0
standard deviation	0.69	1.0
t	2.49	
p	0.029*	

Procedure

After a brief explanation of the experiment, the experimenter began the experiment. Two sessions were conducted during the experiment. The experimenter randomly decided which ECA, DEEP or the gradual method, was used for the first session, and the other ECA was used for the second session. The participant repeatedly selected proposals provided by the ECA until he/she was satisfied his/her end goal for the robot. At the conclusion of each session, the participant completed a questionnaire regarding the ECA ' s evaluations.

Results of accuracy in DEEP

We randomly picked seven participants before the experiment. These seven participants chose their top three emphasizing points out of 23 factors at the end of both session. The reason why we picked up a limited number of participants is that the choice of emphasizing points was very time consuming process because they had to understand the meanings of 23 factors and reflect on their decision-making. Therefore, we could gather a limited number of participants for the research. We then calculated concordance rates between the factors chosen by the user and the factors estimated by each ECA. We conducted a t-test to compare the concordance rates of DEEP with that of the gradual method. Results are shown in Table 1. Average values show the average number of matched factors.

Table 2: Chi-squared for the effect of method on dynamic changes

	changed	not changed
proposed	25	1
gradual	22	4
p	0.158	

Table 3: Sign-test for comparison of ECA method

	score (proposed > gradual)
average	1.0
standard deviation	1.9
p	0.013*

The results of a t-test confirmed that DEEP more accurately estimates emphasizing points than does the gradual method. We suggest that DEEP has sufficient performance for estimating emphasizing points because the average is high and the standard deviation is low. Therefore, by using verbal reactions, body movements and physiological indices, DEEP can correctly estimate the emphasizing points of each participant.

Questionnaire results

The participants answered three rating questions on the ECA's behavior using a seven-point scale. The scale was presented as seven ticks on a black line without numbers, which we scored from -3 to +3.

Each of the three questionnaires contained two kinds of questions; one was on how much the ECA affected participant's thought ("how much" question), another was regarding which method had more affected participant's thought ("which" question).

Changing emphasizing points and purpose of robot Participants answered whether they dynamically changed their emphasizing points and purpose of the robot throughout the interaction ("how much" questions). We performed Chi-squared test to confirm that there was a significant difference between DEEP and the gradual method, and the results are presented in Table 2. Participants also answered which method caused more dynamic changes ("which" question); we performed sign-test to calculate the difference between the two methods, which is shown in Table 3 (when the gradual method caused most changes: -3 - when DEEP caused most changes: +3).

There is no significant difference between the "how much" scores, because both methods could cause dynamic changes during the interaction. This means that humans easily change their emphasizing points even when simple algorithms provide the proposal and explanation. Meanwhile, DEEP caused significantly more changes than did the gradual method. It is possible that participants pay attention to broader factors than contained in the mobile robot task because the proposed method was sensitive to changes in emphasizing points and modified subsequent proposals accordingly.

Table 4: Wilcoxon signed-rank test results on user satisfaction of ECA's final proposal

	proposed	gradual
average	1.8	0.81
standard deviation	2.3	1.6
z	2.11	
p	0.035*	

Table 5: Sign-test results on which ECA provided the best proposal

	score (proposed > gradual)
average	1.1
standard deviation	2.3
p	0.038*

Participant satisfaction of ECA's final proposal Participants answered how satisfied they were with the ECA's final proposal ("how much" questions). The results of a Wilcoxon signed-rank test are shown in Table 4 (not at all: -3 - very much: +3). Participants also answered which method provided a more satisfactory proposal ("which" question). We performed a sign-test, and the results are shown in Table 5 (satisfy the final proposal of the ECA with gradual method: -3 - satisfy the final proposal of the ECA with DEEP: +3).

Both of Table 4 and Table 5 show that the ECA with DEEP provided a significantly more satisfactory proposal than the ECA with the gradual method. However, it is important to note that the standard deviation for the results of the ECA with DEEP in Table 4 and Table 5 are fairly large. We return to the implications of this result in the discussion.

Naturalness of ECA ' s proposals Participants answered how natural the sequence of proposals was ("how much" questions). We performed a Wilcoxon signed-rank test, and the results are shown in Table 6 (not at all: -3 - very much: +3). Participants also answered which method provided more natural proposals ("which" question). The results of a sign-test are shown in Table 7 (the ECA with gradual method provided most natural proposal: -3 - the ECA with DEEP provided most natural proposal: +3).

Both Table 6 and Table 7 show that the ECA with DEEP provided significantly more natural proposals than the ECA with gradual method. The each content of proposals were the same between the proposed method and gradual method. Therefore, naturalness must be attributed to presentation order and whether the proposals reflected their emphasizing points. The proposed method most likely provided more natural proposals because DEEP could quickly reflect changes in their emphasizing points.

Table 6: Wilcoxon signed-rank test results on naturalness of ECA proposals

	proposed	gradual
average	1.2	0.27
standard deviation	1.8	1.6
z	2.4	
p	0.015*	

Table 7: Sign-test results on which ECA provided more natural proposals

	score (proposed > gradual)
average	0.89
standard deviation	1.7
p	0.027*

Discussion

In this study, we evaluated one method for estimating emphasizing points based on verbal and nonverbal information and physiological indices. As a result, we confirmed that our proposed method improved the accuracy of estimating emphasizing points, has more latitude in changing emphasizin points, is natural, and participants are more satisfied with the final proposal. In addition, we find evidence that people often change their emphasizing points and purpose of the task during the interactive decision-making process.

The proposed method considers changes of emphasizing points. Therefore, the proposed method often provided proposals that included new combinations of factors which the participant did not specially emphasize. One participant reported "I was often surprised at the dynamic changes of the proposals." The surprise sometimes causes uncomfortable feelings so we will have to consider proposal history and provide additional explanations for the change.

The standard deviations in proposed method are relatively large. This means that the effectiveness of the proposed method is different across individuals. One of the reasons was that some participants' demands could not be satisfied by the ECA. In those cases, the ECA did not provide any notification of impossibility or alternatives. In many posssible cases, the ECA with DEEP quickly responded to participants' demands, so, in some impossible cases, the participants who had impossible demands felt disappointed, as would be expected. Future work should include notification capabilities.

Conclusion

In this study, we evaluated whether our proposed method, which estimates dynamic changes of emphasizing points based on verbal reactions, body movements, and physiological indices, is useful for interactive decision-making and for selecting a proposal that satisfies the user's end goal. For this purpose, we conducted an experiment that compared two methods: our method and an existing method that gradually estimates emphasizing points based on participants' proposal choice. As a result, we confirmed that DEEP improved estimation accuracy, user satisfaction, and naturalness of proposals. We propose that interactive decision-making be based on estimation of emphasizing points.

One important issue that should be explored in future work is more clearly define the criteria for noting verbal and nonverbal behavior. Physiological indices are very useful for estimating internal states of human but measuring these indices may not be natural in many cases. In future work, we will try to replace physiological indices with synthetic use of some verbal and nonverbal behaviors.

References

Aydogan, R., & Yolum, P. (2007). Learning consumer preferences using semanticsimilarity. In *Aamas '07: Proceedings of the 6th international joint conference on autonomous agents and multiagent systems* (pp. 1–8). New York, NY, USA: ACM.

Bosma, W., & Andre, E. (2004). Exploiting emotions to disambiguate dialogue acts. In *Proceedings of the 9th international conference on intelligent user interfaces* (pp. 85–92).

Iwaki, M., Arakawa, S., & Kiryu, T. (2008). *Influence on biosignal and working efficiency of sound environment in typewriting* (Tech. Rep.). IEICE technical report. ME and bio cybernetics.

Kitamura, M., Shimohata, S., Sukehiro, T., Ikeno, A., Sakamoto, M., Orihara, I., et al. (2008). *Design and development of dialogue system for laddering search service* (Vol. 108; Tech. Rep.). IEICE technical report. Natural language understanding and models of communication.

Kurata, Y. (2010). Interactive assistance for tour planning. *Spatial Cognition 2010 Lecture Notes in Artificial Intelligence*, *6222*, 289–302.

Lin, T., Omata, M., Hu, W., & Imamiya, A. (2005). Do physiological data relate to traditional usability indexes? In *Proceedings of the 17th australia conference on computer-human interaction* (pp. 1–10).

Miyata, H. (Ed.). (1998). *The new physiological psychology (japanese)* (Vol. 3). Kitaohji-shobo.

Nakazono, K., Hada, T., Ataka, E., Tanaka, H., & Nagashima, Y. (2008). *Workload evaluation of gaming task by physiological indices and psychological indices* (Vol. 107 - 553; Tech. Rep.). Technical report of IEICE. HIP.

Ohmoto, Y., Kataoka, M., Miyake, T., & Nishida, T. (2011). A method to dynamically estimate emphasizing points and degree by using verbal and nonverbal information and physiological indices. In *The 2011 ieee international conference on granular computing 2011* (pp. 508–514).

Prendinger, H., & Ishizuka, M. (2005). The empathic companion: A character-based interface that addresses users'affective states. *Applied Artificial Intelligence*, *19*(3-4), 267–285.