International Journal of Innovation and Research in Educational Sciences
Volume 6, Issue 3, ISSN (Online) : 2349–5219

# Prediction of Students' Answer Relevance in Discussion Based on their Heart-Rate Data

**Shimeng Peng[1], Shigeki Ohira[2] and Katashi Nagao[3*]**
[1] Graduate School of Informatics, Nagoya University, Nagoya, Japan.
[2] Information Technology Center, Nagoya University, Nagoya, Japan.
[3*] Graduate School of Informatics, Nagoya University, Nagoya, Japan.
*Corresponding author email id: nagao@i.nagoya-u.ac.jp

*Date of publication (dd/mm/yyyy): 15/06/2019*

***Abstract* – Whether a discussion is executed effectively depends on the completion level of the question-and-answer segments (Q&A segments) generated during the discussion. Relevance of answers could be used as a clue for evaluating the Q&A segments' completion degree. In this study, we argue that discussion participants' heart rate (HR) and its variability (HRV), which have recently received increased attention for being a crucial indicator in cognitive task performance evaluation, can be used to predict participants' answer-relevance in Q&A segments of discussions. To validate our argument, we propose an intelligent system that acquires and visualizes the HR data with the help of a non-invasive device, e.g. an Apple Watch, for measuring and recording the HR data of participants which is being updated in real-time. We also developed a web-based human-scoring method for evaluating answer-relevance of Q&A segments and question-difficulty level. A total of 17 real lab-seminar-style discussion experiments were conducted, during which the Q&A segments and the HR of participants were recorded using our proposed system. We then experimented with three machine-learning classifiers, i.e. logistic regression, support vector machine, and random forest, to predict answer-relevance of Q&A segments using the extracted HR and HRV features. Area under the ROC Curve (AUC) was used to evaluate classifier accuracy using leave-one-student-out cross validation. We achieved an AUC = 0.76 for logistic regression classifier, an AUC = 0.77 for SVM classifier, and an AUC = 0.79 for random forest classifier. We examined possibilities of using participants' HR data to predict their answer-statements' relevance in Q&A segments of discussions, which provides evidence of the potential utility of the presented tools in scaling-up analysis of this type to a large number of subjects and in implementing these tools to evaluate and improve discussion outcomes in higher education environment.**

***Keywords* – Answer Relevance Prediction, Learning Analytics, Discussion Mining, Machine Learning, Heart Rate Variability.**

## I. INTRODUCTION

Seminar-style discussion, one of the most familiar types of intellectual and creative activities held regularly at university laboratories, is where research ideas are generated and exchanged. Usually, such a discussion consists of one presenter presenting his or her current research progress, with many questions being raised to challenge and point out problems around the presentation content by participants often involving peers and instructors, and answers by the presenter. We call question and answer pairs Q&A segments. Relevant answers in discussions help to inspire new research ideas and grasp the research direction as well as improve the discussion outcomes. To that end, it is necessary to scientifically evaluate and predict how relevant are students' answers in Q&A segments during discussions.

In this study, we generate an intelligent system that acquires and visualizes discussion participants' heart rate (HR) data during lab-seminar-style discussion experiments and analyze how this data can be used to predict answer-relevance while students complete the Q&A sessions in discussion activities.

## II. Related Work and Motivation

A substantial amount of prior work on improving discussion outcomes mainly focuses on evaluating and training students' discussion abilities; these abilities were evaluated using audio and video cues such as "voice quality", "speech speed", "pause", and "conciseness". In other words, students' ability to speak in an easy-to-understand way was evaluated [1] [2]. With the introduction of 3D sensing devices such as Microsoft Kinect, tracking students' movements during discussions and presentations as cues for discussion-skill evaluation has increased [3] [4]. Chen and Feng's research [5] combined audio, video, and human-movement information as multimodal cues and generated an automatic assessment system that establishes a set of specific body-movement standards, such as hand gestures and eye direction, to evaluate students' discussion skills quantitatively with the aim of improving discussion outcomes.

While studies have been carried out on evaluating students' discussion skills based on acoustics and gesture information, little attention has been paid to the relevance of discussion statements, especially the relevance of answer statements in discussion activities. In an academic seminar-style discussion environment, presenters are expected to answer within a few seconds and as accurately as possible in front of the questioners, which requires high level of question comprehension, strong communication skills, and improved mental control, all representing high discussion ability. In addition, many appropriate answers given by presenters could generate more inspiring questions and discussion enthusiasm in participants.

However, previous answer-relevance evaluation or prediction studies often used application scenarios involving social question-and-answer sites (Q&A sites) such as "Yahoo! Answers" and "Answer.com", on which the answerers have enough time to think of and repeatedly modify their answers. In addition, there was no face-to-face interaction between the questioners and answerers. Based on these specific conditions, these studies used natural-language-processing tasks. Belinkov et al. [6] computed text-based features based on yes-like words in the answer statements. Patil and Lee [7] analyzed certain linguistic features to identify expert answers. Effectively evaluating students' answer-relevance in discussion activities is challenging; therefore, our goal is to find a more accurate method to evaluate answer-relevance in Q&A segments generated by the on-the-spot discussion scenarios.

Taking into account that the discussion process is a classical type of cognitive activity which could elicit a subjective state response with some changes in certain physiological data, such as HR and its variability (HRV), several studies have proven that HR is an important index of the autonomic nervous system's regulation of the cardiovascular system [8] [9] [10]. Therefore, there has been an increasing focus on observing the correlation between the HR data and cognitive activity performance. Hellhammer and Schubert [11] assessed HR changes before, during, and after Trier Social Stress Test (TSST) including a series of cognitive tasks. Muthukrishnan et al. used HRV to predict human-cognitive performance [12]. We decided to take advantage of this theoretical background to estimate the subjective state of discussion participants by measuring their HR and HRV data during discussions and attempted to evaluate their answer-relevance in Q&A segments using such data.

There are several novel aspects in which this study differs from prior related work. (1) Instead of evaluating discussion participants' acoustics and gesture manners, we are interested in challenging the more complex work on answer-statements' relevance prediction. Regarding randomness, openness, and improvisation of the questions raised in discussion activities, as well as subjective influences of the answer-relevance on the

evaluation, we set up a web-based human scoring system to collect timely evaluations of answer-relevance from all discussion participants, including the questioners themselves, which ensures scientific reliability of the research results. (2) Different from other existing studies which analyzed semantic information of Q&A segment's statements to evaluate answer-statements' relevance, we put our research emphasis on the interactive face-to-face conversation-style Q&A segments in discussions and suggest to use presenter's physiology data to predict the relevance of the answer-statements. (3) The subjects in this study are university students who are holding regular lab-seminar-style discussion activities, in which one student gives a presentation around his or her recent research progress and other participants raise questions around the current presentation content. Unlike in our research, other studies typically consider a specific cognitive task such as TSST [11]. We hope to explore how physiology data, i.e. HR, could be used to predict answer-relevance of Q&A segments in a non-baseline cognitive task.

## III. Discussion Experimental Design and Data Acquisition

### *Discussion Experimental Scenario and Participants*

We conducted discussion experiments based on a real lab-seminar-style discussion environment in which a presenter explains a research topic while displaying slides, and Q&A session between the presenter and the meeting participants is conducted during the presentation. A semi-automatic discussion system called discussion mining (DM) system [13] was developed and used in our laboratory to record the content of face-to-face discussion experiments while providing metadata. As shown in Figure 1, the DM system using multiple cameras and microphones is installed in a discussion room to record the discussion experimental scenario and the detailed content of Q&A statements which are segmented based on the start and end time of each speech. In the centre of the discussion room, there is also a main screen that displays presentation materials and demonstration videos, and on both sides, there are sub-screens for displaying the information and images of the currently speaking participants. A secretary manually records the contents of the Q&A conversation, and each speaker tags his/her speech. Therefore, the Q&A segment data used in this study reproduces real Q&A conversations happened in face-to-face discussion activities. Each discussion experiment was held once a week for 1.5-2 hours, during which we asked one student to be a presenter and give a presentation on their recent research progress and the others to raise questions as questioners. The 15 participants included four undergraduates, eight graduate students, and three professors; two participants were female students and the rest were male.

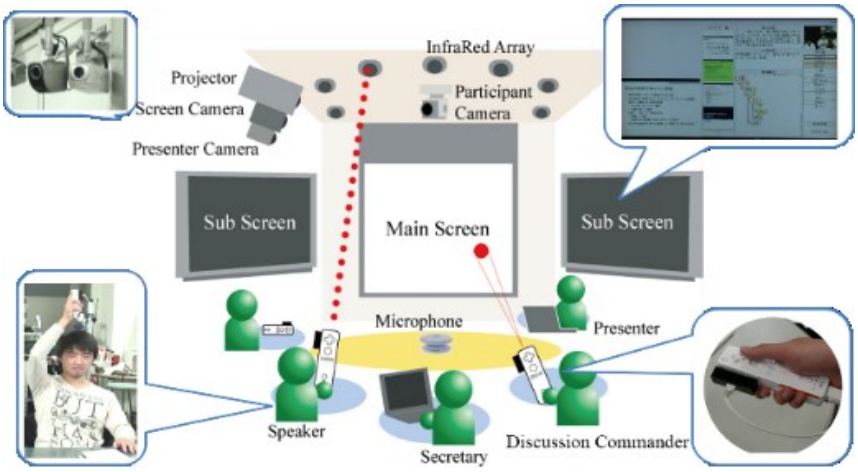


Fig. 1. Discussion experimental scenario recorded by DM system.

*Data-Acquisition System*

We aim to collect and use presenters' heart rate data during discussion experiments to predict answer-relevance in the Q&A segments. In contrast to medical HR measurement devices, such as chest-worn electrocardiogram (ECG) sensors, we used non-invasive wireless smart trackers, such as Apple Watch, Fitbit series, Microsoft Bands, Jawbone Up, and Samsung Galaxy Gear, which could reduce the cost and difficulty of monitoring the HR data in real-world applications, as well as minimize the awareness of the monitoring process during experiments. These devices use light-emitting diodes to measure the frequency at which the blood pumps. El-Amrawy and Nounou [14] compared the accuracy of monitoring the HR data between smart trackers, such as Apple Watch and Samsung Gear, and a professional HR monitoring device and found that Apple Watch showed the highest accuracy and precision. Given the high accuracy of Apple Watch in HR data measuring and its wide range of uses, we adopted Apple Watch Series 2 to collect presenters' HR data and visualize their HR information after discussions.

To collect presenters' HR data in our discussion experiments, we used a real-time HR data acquisition and visualization system developed in our previous research work [15]. In each experiment, one presenter was asked to wear Apple Watch Series 2 on their left hand before each discussion, as shown in Figure 2. He or she could choose their own ID, press the "start" button to start their HR data acquisition, and press "stop" at the end of the discussion. Based on the Health Kit framework, the HR data of the presenter was measured almost in real-time in 5-7 sec intervals, and the collected HR data and presenter's information were displayed on the Apple Watch screen, as well as on the HR browser.


Fig. 2: Presenter's HR data acquisition.

The HR browser that we developed provides an intuitive method for discussion presenters to check and observe their HR information during the discussion experiments. All HR data collected from our HR data acquisition system were presented synchronously in the heart-rate browser with a time line of the corresponding discussion and managed by the discussion participants. As shown in Figure 3, there are three components in the HR browser, i.e. a search menu, an HR graph, and the HR records.

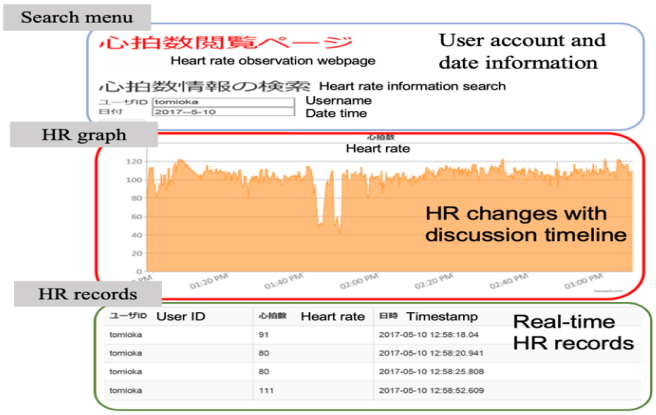



Fig. 3: Discussion presenters' HR browser.

(1) Search menu: Historical HR data and user information can be searched through this menu at the top of the browser.

(2) HR graph: The graph provides an intuitive way to observe changes in presenter's HR data throughout the discussion.

(3) HR records: The HR data at each point of the discussion with the presenter's information can be checked here.

For the purpose of collecting the relevance-evaluation scores of Q&A segments from discussion participants as the "ground truth" labels, we developed a webpage-based human-scoring method with our DM system, which was launched on tablets hold by participants during experiments. Each time the DM system recognized that the presenter has answered a question, the tablets would automatically switch to the Q&A-segment evaluation page, as shown in Figure 4. All discussion participants including those asking a question were required to give timely evaluation scores of not only the answer-relevance but also of the question difficulty.

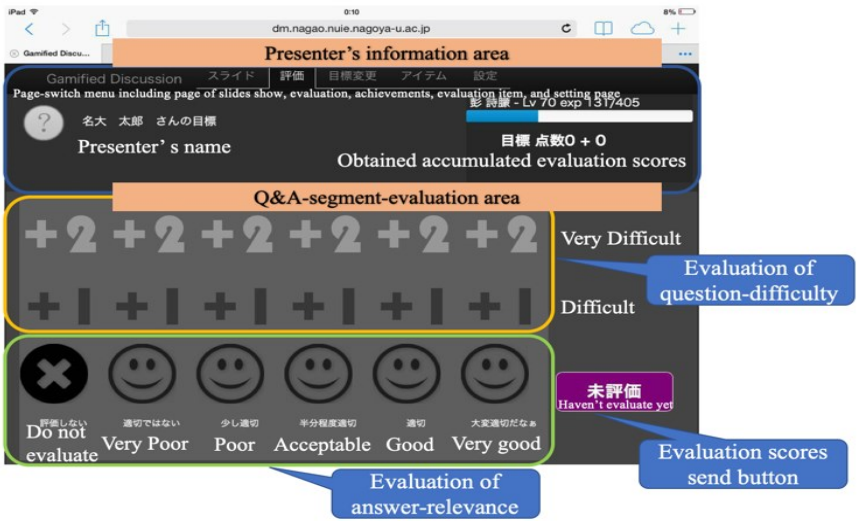



Fig. 4. Webpage-based human-scoring method.

There are two main areas in this web evaluation page. (A) Presenter's information area: The current presenter's name is shown on the top-left and the obtained accumulated evaluation scores are shown on the top-right of the page. (B) Q&A-segment evaluation area: In this area, all participants need to give two evaluation scores after each Q&A segment. One score is given for answer-relevance on a five-point scale from very poor to very good by tapping different "smile" buttons at the bottom of the page. Participants also had to evaluate the difficulty of the question in the same Q&A segment by tapping "+1" (difficult) or "+2" (very difficult) on top of each smile button. Otherwise, the default evaluation value was "Simple". After all evaluation actions were completed, participants had to tap the "send" button, so that the evaluation scores could be sent to the database and shown in the "Accumulated evaluation scores area" on the top-right of the tablet page. There is also an item "Do not evaluate" if a discussion statements was not a question but a comment.

## IV. Answer-Relevance Prediction Methodology

We conducted a total of 17 real lab-seminar-style discussion experiments, in which one student gave a presentation and answered questions from other participants according to the talk content. Among 15 experiment participants (only students would give presentations), four second-year graduate students and one first-year

graduate student gave presentations twice; the remaining students gave presentations once. Finally, 247 Q&A segments with a mean length of 5 minutes per Q&A segment were generated and recorded; 14 to 15 Q&A segments with a presenter and participants were conducted during each discussion experiment. The Q&A segments' answer-relevance and difficulty level of questions were evaluated using our webpage-based human-scoring method by all participants after each question was answered. We acquired the entire HR data of presenters in each discussion experiment.

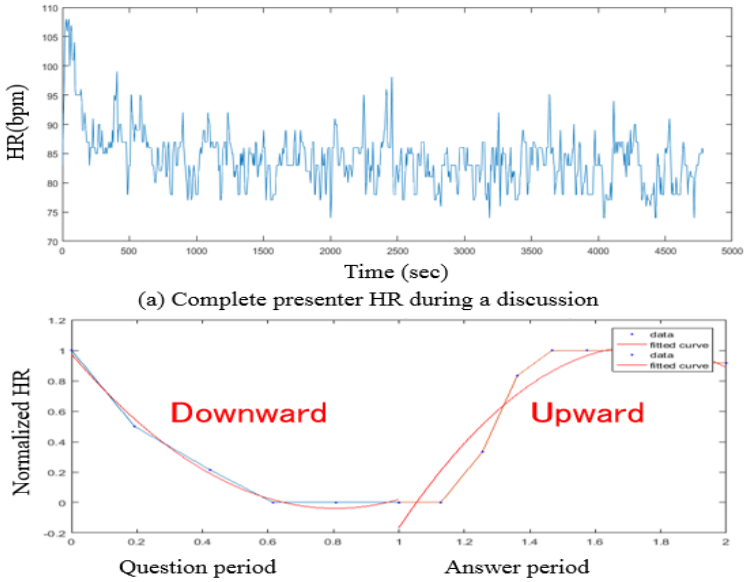
### *Qualitative Analysis*

A total of 247 Q&A segments were generated from 17 discussion experiments. Our starting point was to categorize the answers in Q&A segments into relevant or irrelevant depending on how correctly the presenter answered the questions. Therefore, after each question was answered, all discussion participants were asked to score the relevance of the answer on a five-point scale: 1 = very poor, 2 = poor, 3 = acceptable, 4 = good, 5 = very good. The answers were considered relevant if the scores were 4–5 and irrelevant if the scores were 1–3. As a result, 112 Q&A segments were evaluated as relevant and 135 Q&A segments were evaluated as irrelevant. We then tried to investigate whether there is a difference in evaluation scores between the participant who asked a question and other discussion participants. We used Kappa coefficient to measure the inter-rater agreement between different items. If the value of Kappa varies from 0.41 to 0.6, the agreement level is considered moderate, and if it is within the range of 0.6–0.8, the agreement between different subjective opinions is considered substantive. If Kappa is in the range of 0.81–0.99, two participants can be considered as having reached almost perfect agreement [16]. We obtained a Kappa of 0.67, which means the questioners of the Q&A segments had substantively the same evaluation opinions regarding answer-relevance with the other participants in the same discussion. These results indicate that in our future work, we can take into account only the evaluation opinions of questioners, and this will decrease the work load on other participants.

To investigate if the difficulty of questions affects the relevance of the answers, we also recorded the difficulty level of questions evaluated by participants other than questioners, because we believe that the perspective of a third party is more objective and can be referenced. Among all Q&A data, only seven questions in Q&A segments were evaluated as "Difficult". No questions were evaluated as "Very difficult". Therefore, we assume that not many difficult or very difficult questions would come up in a lab discussion at a university. This makes us believe that there is no need to consider question difficulty level in Q&A segments at this stage. In our study, we did not classify the difficulty of the Q&A segments' questions in advance, as they would not affect the experimental results.

### *Quantitative Analysis*

We aim to determine if the HR data of the presenters can be used to predict their Q&A segments' answer-relevance. For that purpose, the HR data of the presenters were analyzed based on their Q&A segments generated in the discussions. All HR information of the presenter during one discussion experiment, which lasted for almost 1.5 hours, is shown in Figure 5 (a). The horizontal axis represents the relative time-line of a discussion, and the vertical axis represents the change in the corresponding HR. The HR data were then extracted and segmented with each corresponding Q&A segment. Figure 5 (b) shows the presenter's HR data in three periods: during question period (blue line), during answer period (orange line), and during both periods.




(a) Complete presenter HR during a discussion

(b) HR trend during question and answer period

Fig. 5. HR data of presenter in discussion experiment.

We computed 18 HR and HRV features, including mean, standard deviation (std.), and root mean square successive difference (RMSSD), from all Q&A segments, as well as from the separate question and answer periods. These features have been proven as important for understanding the HRV differences under cognitive activities [17]. The trends in the HR of these three periods were also computed by calculating the difference between two adjacent HR points. If the number of positive differences was more than the negative ones, we assumed that the HR period shows an upward trend; otherwise, the HR period shows a downward trend (Figure 5 (b)). We used a quadratic curve (red line) to demonstrate the HR trend more clearly for the readers. We can see that the HR shows a downward trend during the question period and an upward trend during the answer period. We also divided the HR data of these three periods into the following nine ranges: less than 60, 60–70, 71–80, 81–90, 91–100, 101–110, 111–120, 121–130, and more than 130 beats per minute (bpm). The mean and std. were calculated to describe the HR appearance-frequency distribution in each range. These metrics in the frequency domain during the question period, answer period, and Q&A segments were calculated separately. As a result, we obtained 18 HR and HR features including All_mean, All_std., All_rMSSD, All_trend Freq_all_mean, Freq_all_std., Que_mean, Que_std., Que_rMSSD, Que_trend, Freq_que_mean, Freq_que_std., Ans_mean, Ans_std., Ans _rMSSD, Ans _trend, Freq_ ans _mean, and Freq_ans_std.

*Predictive modeling of answer-relevance in Q&A segments*

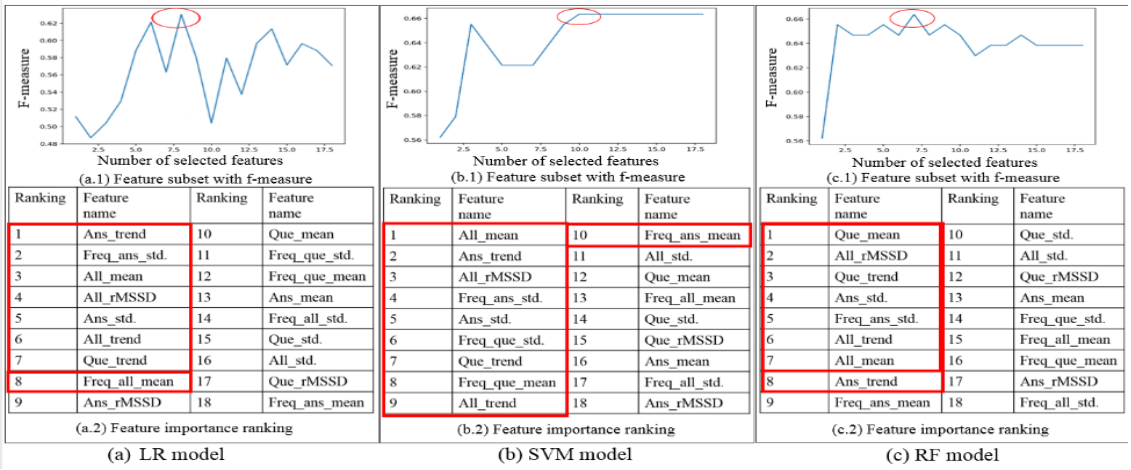
To explore whether the HR data of presenters could be used to predict their Q&A segments' answer-relevance during discussions, we experimented with different types of machine-learning predictive classifiers, including logistic regression, support vector machine, and random forest. We performed leave-one-student-out cross validation and reported the mean area-under-curve (AUC) scores with 95% confidence intervals for each predictive classification model. In addition, a feature selection technique was applied to all HR and HRV features (only on training data, not testing data) in order to reduce the dimensionality of the feature space.

It is possible to use all 18 HR and HRV features for these three predictive classifiers; however, this may decrease the performance of the classifiers, particularly because of dimensionality. Therefore, we attempted to find a subset of HR and HRV features that could be used to discriminate the relevant and irrelevant answer

classes with the highest classification accuracy, which indicates the highest F-measure. Therefore, on our training data set, we used recursive feature elimination [18] with cross validation (RFECV), which ranks the features according to their importance to the different classifiers, and determine the best size of the feature subset. By calculating the F-measure (F1 score), which is the harmonic mean of precision and recall, we extracted the feature subset that could achieve the best evaluation performance for the corresponding classifiers.

*Results and Discussion*

Figure 6 shows three sub-graphs that illustrate separately the best subsets of all HR and HRV features for each classifier at the top and the feature-importance-ranking results at the bottom. The highest F-measure was obtained when eight or seven key features were included in the subsets for the LR classifier; therefore, the first seven or eight features surrounded by the red rectangle were considered as a feature subset for the LR model. Similarly, the first ten features comprised the candidate subset for the SVM classifier, and there were two candidate feature subsets for the RF-based model, which involved seven or eight features in the ranking list counted from the top.




| Ranking | Feature name | Ranking | Feature name |
|---|---|---|---|
| 1 | Ans_trend | 10 | Que_mean |
| 2 | Freq_ans_std. | 11 | Freq_que_std. |
| 3 | All_mean | 12 | Freq_que_mean |
| 4 | All_rMSSD | 13 | Ans_mean |
| 5 | Ans_std. | 14 | Freq_all_std. |
| 6 | All_trend | 15 | Que_std. |
| 7 | Que_trend | 16 | All_std. |
| 8 | Freq_all_mean | 17 | Que_rMSSD |
| 9 | Ans_rMSSD | 18 | Freq_ans_mean |

(a.2) Feature importance ranking

(a) LR model

| Ranking | Feature name | Ranking | Feature name |
|---|---|---|---|
| 1 | All_mean | 10 | Freq_ans_mean |
| 2 | Ans_trend | 11 | All_std. |
| 3 | All_rMSSD | 12 | Que_mean |
| 4 | Freq_ans_std. | 13 | Freq_all_mean |
| 5 | Ans_std. | 14 | Que_std. |
| 6 | Freq_que_std. | 15 | Que_rMSSD |
| 7 | Que_trend | 16 | Ans_mean |
| 8 | Freq_que_mean | 17 | Freq_all_std. |
| 9 | All_trend | 18 | Ans_rMSSD |

(b.2) Feature importance ranking

(b) SVM model

| Ranking | Feature name | Ranking | Feature name |
|---|---|---|---|
| 1 | Que_mean | 10 | Que_std. |
| 2 | All_rMSSD | 11 | All_std. |
| 3 | Que_trend | 12 | Que_rMSSD |
| 4 | Ans_std. | 13 | Ans_mean |
| 5 | Freq_ans_std. | 14 | Freq_que_std. |
| 6 | All_trend | 15 | Freq_all_mean |
| 7 | All_mean | 16 | Freq_que_mean |
| 8 | Ans_trend | 17 | Ans_rMSSD |
| 9 | Freq_ans_mean | 18 | Freq_all_std. |

(c.2) Feature importance ranking

(c) RF model

Fig. 6. HR and HRV feature selection for each classifier.

We then generated three classifiers using the selected seven HR and HRV features involving all_mean, answer_trend, all_RMSSD, freq_answer_std., answer_std., question_trend, and all_trend, which exhibited the largest effect on all three classifiers. We reported mean AUC scores to evaluate our classifiers' overall prediction ability. In addition to mean AUC scores, we reported mean recall scores on the irrelevant answer prediction task, which refers to the percentage of total irrelevant answers correctly classified by classifiers, as shown in Table 1.

Table 1. AUC scores of classifiers and mean recall scores of irrelevant answer recognition task

| Model | Mean AUC | Mean recall scores on irrelevante answer recognition task |
|---|---|---|
| LR | 0.76 | 0.69 |
| SVM | 0.77 | 0.65 |
| RF | 0.79 | 0.78 |

As the results reported in Table 1 show, for the Q&A segments' answer-relevance prediction task, we got good mean AUC scores for all classifiers, especially for the RF classification model. This suggests that presenters' heart rate data could be used to effectively predict their answer-relevance in lab-style discussion

experiments. In addition, we reported mean recall scores based on irrelevant answer recognition task, for the same student's Q&A segments; the RF classifier could recognize irrelevant answers better than the other two classifiers. Furthermore, the LR classifier presents a stronger identification ability than the SVM classifier in the irrelevant answer recognition task, even though it shows a slightly lower overall classification performance.

Our ultimate goal is to improve discussion outcomes in higher education environment. We hope that university students will be able to give more relevant answers in the Q&A sessions during discussion activities. With our proposed methodology, we can help students to effectively recognize their irrelevant answers in lab discussions. Furthermore, we would like to understand the reasons why they cannot give correct answers in discussions: is it because the questions are too difficult, or is it because the students have not prepared well for the discussion, or something else?

We discussed in the previous subsection whether we should classify the difficulty level of questions in advance. Although the results show that there is no need to consider the difficulty level, it is possible that we did not collect enough data to support our analysis in evaluating the answering performance of presenters at different difficulty levels in the Q&A segments. Therefore, we will continue to use our webpage-based human-scoring method to collect the Q&A segment data of different difficulty levels for future research and will carry out further analysis. In our leave-one-student-out cross-validation evaluation process, we also looked at the classification results of each student. Master students who have more experience in discussion presentations showed more false negative recognition samples. This means that the answers evaluated by discussion participants as relevant were recognized by the classifiers as irrelevant because of the presenters' low-confidence mental state when giving answers. We analysed the statement content of these Q&A segments and conducted interviews with the master students who answered these Q&A. We found that in their answers, there were often words expressing uncertainty, such as approximate, rough, maybe, and perhaps. The following example demonstrates our findings. A presenter was evaluating the performance of a tennis player's serving style. A participant raised a question: "How did you determine the threshold of serving time?", to which the presenter-student answered: "I roughly estimated the value based on the approximate time of the batting". This example shows that from the perspective of the questioner, the presenter correctly answered the question using the available language skills, but from the answer content, the presenter clearly lacked self-confidence because his answer revealed an apparent inadequacy in his research. This issue should be addressed in our future work.

## V. Conclusion and Future Work

In this study, we argued that students' physiological data, such as HR, can be used to effectively predict their answer-relevance in Q&A segments during discussions. To provide sufficient experimental evidence in support of our argument, we conducted 17 real lab-seminar-style discussion experiments, in which one lab student gave a presentation and other participants asked one or two questions about the content of the presentation. We recorded 247 Q&A segments from discussion experiments in details. A real-time HR data acquisition and visualization system was developed to collect presenters' HR data with the help of a non-invasive device, i.e. Apple Watch, during the discussion experiments. A real-time webpage-based human-scoring method was designed to collect the relevance evaluation scores of Q&A segments from discussion participants as the "ground truth" labels.

Three binary classification machine-learning methods, logistic regression, support vector machine, and rando-

-m forest, were adopted to construct predictive classifiers for the Q&A segments' answer-relevance prediction task. We also selected seven HR and HRV features that had significant meaning to all classifiers. Regarding our models' overall classification performance, a mean AUC of 0.76 for the LR-based model, 0.77 for the SVM-based model, and 0.79 for the RF-based model were obtained. We also observed mean recall scores of each classifier related to the irrelevant answer identification task. The results confirmed our hypothesis that the HR data of the presenters could be used to effectively predict their answer-relevance in discussion activities. Furthermore, our findings collectively provide initial evidence for the possibilities of using students' physiological data to evaluate their discussion activity performance for the purpose of improving the discussion outcomes.

In our future work, we will focus on the application of the argument brought forward in this paper to improve students' discussion outcomes. We plan to develop a follow-up function, in which feedback regarding irrelevant Q&A segments is given to presenters after discussions to encourage them to spend more time on comprehending the questions, to sort out their research in order to find more relevant answers, to strengthen their communication skills to be able to give participants a more understandable description, and in the long run, to effectively improve students' discussion outcomes.

## REFERENCES

[1] K. Kurihara, M. Goto, J. Ogata, Y. Matsusaka, and T. Igarashi, "Presentation sensei: a presentation training system using speech and image processing," in *Proceedings of the 9th International Conference on Multimodal Interfaces*, 2007, pp. 358–365.
[2] S. Scherer, G. Layher, J. Kane, H. Neumann, and N. Campbell, "An audiovisual political speech analysis incorporating eye-tracking and perception data," in *LREC*, 2012, pp. 1114–1120.
[3] Z. Zhang, "Microsoft Kinect sensor and its effect," *IEEE Multi-Media,* 2012, 19(2), pp. 4–10.
[4] M. Swift, G. Ferguson, L. Galescu, Y. Chu, C. Harman, H. Jung, I. Perera, Y. C. Song, J. Allen, and H. Kautz, "A multimodal corpus for integrated language and action," in *Proceedings of the International Workshop on Multi Modal Corpora for Machine Learning*, 2012.
[5] L. Chen, G. Feng, C.W. Leong, J. Joe, C. Kitchen, and C.M. Lee, "Designing an automated assessment of public speaking skills using multimodal cues," *Journal of Learning Analytics*, 2016, 3(2), pp. 261–281.
[6] Y. Belinkov, M. Mohtarami, S. Cyphers, and J. Glass, "Vector SLU: A continuous word vector approach to answer selection in community question answering systems," in *Proceedings of the 9th International Workshop on Semantic Evaluation (Sem Eval 2015)*, pp. 282–287.
[7] S. Patil and K. Lee, "Detecting experts on Quora: by their activity, quality of answers, linguistic characteristics and temporal behaviors," *Social Network Analysis and Mining*, 2016, 6(1), p. 5.
[8] M.N. Levy, P.J. Schwartz, and K.P. Anderson, *Vagal Control of the Heart: Experimental Basis and Clinical Implications*. Armonk, NY Futura Publishing Company, 1994.
[9] A.J. Camm, M. Malik, J.T. Bigger, G. Breithardt, S. Cerutti, R.J. Cohen, P. Coumel, E.L. Fallen, H.L. Kennedy, R.E. Kleiger, and F. Lombardi, "Heart rate variability: standards of measurement, physiological interpretation, and clinical use," *Circulation*, 1996, 93, pp. 1043–1065.
[10] U.R. Acharya, K.P. Joseph, N. Kannathal, C.M. Lim, and J.S. Suri, "Heart rate variability: a review," *Medical and Biological Engineering and Computing,* 2006, 44(12), pp.1031–1051.
[11] J. Hellhammer and M. Schubert, "The physiological response to trier social stress test relates to subjective measures of stress during but not before or after the test," *Psychoneuroendocrinology,* 2012, 37(1), pp. 119–124.
[12] S.P. Muthukrishnan, J.P. Gurja, and R. Sharma, "Does heart rate variability predict human cognitive performance at higher memory loads?," *Indian Journal of Physiology and Pharmacology,* 2017, 61(1), pp. 14–22.
[13] K. Nagao, K. Kaji, D. Yamamoto, and H. Tomobe, "Discussion mining: Annotation-based knowledge discovery from real world activities," in *Proceedings of the Pacific-Rim Conference on Multimedia*, 2004, Springer, pp. 522–531.
[14] F. El-Amrawy and M. I. Nounou, "Are currently available wearable devices for activity tracking and heart rate monitoring accurate, precise, and medically beneficial?," *Healthcare Informatics Research,* 2015, 21(4), pp. 315–320.
[15] S. Peng and K. Nagao, "Automatic evaluation of presenters' discussion performance based on their heart rate," in *Proceedings of the 10th International Conference on Computer Supported Education (CSEDU 2018)*, 2018, pp. 27–34.
[16] J. Landis and G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, 1997, 33(1), pp. 159–174.
[17] X. Wang, X. Ding, S. Su, Z. Li, H. Riese, J. F. Thayer, F. Treiber, and H. Snieder, "Genetic influences on heart rate variability at rest and during stress," *Psychophysiology*, 2009, 46(3), pp. 458–465.
[18] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning,* 2002, 46(1-3), pp. 389–422.

## AUTHORS PROFILE'

**Shimeng Peng**
Shimeng Peng is a second year of Ph.D. student at the Department of Intelligent Systems, Graduate School of Informatics, Nagoya University. Her research interests are analyzing multimodal wearable sensory signals such as physiological data, facial expressions during student learning activities and applying machine learning methods to predict students' affective state such as nervous, engagement, interest and their impact on learning performance in order to guide improvements of their learning outcomes. email id: hou@nagao.nuie.nagoya-u.ac.jp

**Shigeki Ohira**
Shigeki Ohira is an assistant professor in the Information Media Division of Information Technology Center, Nagoya University. His research interests include educational technology, multimedia content processing and real-world computing. Mr. Ohira has an MIS from the Graduate School of Science and Engineering at Waseda University.
email id: ohira@nagoya-u.jp

**Katashi Nagao**
Katashi Nagao received his B.E., M.E., and Ph.D. in Computer Science from Tokyo Institute of Technology in 1985, 1987 and 1994, respectively. Since 1987, he has been researching natural language processing and machine translation systems at IBM Research, Tokyo Research Laboratory. In 1991, he began conducting research projects on natural language dialogue, multiagent systems, and human-computer interaction at Sony Computer Science Laboratories, Inc. From 1996 to 1997, he was a visiting scientist at the Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, USA. He re-joined IBM's Tokyo Research Laboratory and launched the Semantic Transcoding Proje--ct in 1999. He joined Nagoya University as an associate professor at the Graduate School of Engineering in 2001. Since 2002, he has been researching artificial intelligence and computer-assisted education as a professor at the Graduate School of Information Science, Nagoya University. The Graduate School of Information Science was reorganized into the Graduate School of Informatics in 2017.
email id: nagao@i.nagoya-u.ac.jp