Video Feedback System for Teaching Improvement Using Students' Sequential and Overall Teaching Evaluations

Yusuke Kometani[1], Takahito Tomoto[2], Takehiro Furuta[3], and Takako Akakura[2]

[1] Graduate School of Engineering
[2] Faculty of Engineering, Tokyo University of Science,
1-3 Kagurazaka, Shinjuku-ku,Tokyo 162-8601 Japan
{kometani,tomoto,akakura}@ms.kagu.tus.ac.jp
[3] Nara University of Education, Takahata-cho, Nara-city, Nara 630-8528, Japan
takef@nara-edu.ac.jp

Abstract. We propose a system that allows university teachers to check the effectiveness of their lecture videos and to grasp points for improvement in the lectures. The system offers two functions: time-series graphing, which visualizes real-time changes in students' evaluation during a lecture, and teaching-behavior estimation, which shows teachers information on their own teaching behaviors estimated from the overall evaluation by students of a lecture. The system was developed and evaluation experiments of each function were conducted. The subjective evaluation of each function by teachers showed the following: (1) the time series graph function was useful to narrow down which portion of the lecture videos contained points for improvement and (2) the teaching behavior estimation function was useful to determine the tendency of teaching behavior in a lecture.

Keywords: Sequential evaluation, Overall evaluation, Teaching improvement, Lecture video, Teaching behavior, Student evaluation.

1 Introduction

We present here a methodology to support university teachers to achieve improvements in lecturing. Our approach is to provide a web-based system that allows teachers to review their own lecture videos. However, teachers often need support to grasp the exact points for improvement in their lectures, thus we propose a system that presents lecture evaluation data obtained from the students in conjunction with a lecture video to make teachers aware of areas for improvement.

The evaluations obtained from students can be roughly divided into the following three timeframes: (1) end-of-term, (2) each lecture, and (3) real-time during a lecture. Here we focus on items (2) and (3), which we refer to as "overall evaluation" and "sequential evaluation," respectively. We have designed a feedback system for teachers that provides lecture videos along with sequential and overall evaluation information. The purpose of this study was to verify the usefulness of this system as a means of support for teachers seeking how their lectures may be improved.

S. Yamamoto (Ed.): HIMI/HCII 2013, Part III, LNCS 8018, pp. 79–88, 2013.

The problems that may arise when a teacher develops a strategy to improve a lecture can be outlined as follows. First, to identify areas to improve, time is required to look through the lecture video. The teacher must check the entire video, including portions of it where information is not directly relevant to improvements. We have therefore designed a function that presents a summary of the sequential evaluation data in order to allow the teacher to narrow down in advance the areas that should be improved. Second, teachers will most likely look for which of their teaching behaviors need to be improved. However, monitoring teaching behavior by watching individual lecture videos is not efficient. To solve this problem, we have constructed a model to estimate the behavior of a teacher from an overall evaluation of each lecture. We have designed and incorporated into the system a function to estimate teaching behavior based on the results of the constructed model.

We formulated the following research questions based on the above considerations:

1. Is the sequential evaluation function useful in aiding teachers to identify areas in the lecture video that need improvement?
2. Which teaching behavior estimation models are applicable to multiple lectures on different topics?
3. Can teaching behavior estimation based on overall evaluations help teachers gain insight into how to improve their lectures?

2 Related Studies

Lecture evaluation feedback methods and student reactions to support lecture improvement have been studied in the past. Stalmeijer et al. explored whether feedback effectiveness improved when physician teachers' self-assessments were added to written feedback based on student ratings[3]. The physician teachers considered the combination of self-assessment and student ratings more effective than either self-assessment or written feedback alone. The authors concluded that self-assessment can be useful in stimulating teaching improvement. However, there was no evidence that the teachers grasped points for improvement by reviewing individual lecture videos. Thus, our proposed method additionally involves an objective evaluation of the teaching behavior based on evaluations obtained from the students and the data are then fed back to the teacher.

Hanakawa and Obana showed that lecture improvement was possible through the use of a Twitter log[4]. They developed a system that supported student-teacher interaction during large-scale lectures. During lectures, this system allowed students to send the teacher their reactions and messages, through a handheld unit, related to the clarity of the lesson. A common pattern between lectures was discovered by analyzing the evaluation log and the students' reactions. The teacher was able to know the points of improvement by comparing the lecture video against the common pattern identified. In addition, after teachers strived to improve their teaching, negative evaluations from the students decreased. Hanakawa and Obana's method identifies a common pattern that emerges across multiple lectures, whereas the present study

suggests a method to help teachers understand areas for improvement when watching an individual lecture video.

3 System Design

Different lecture forms, shown in Table 1, were considered in the process of acquiring sequential and overall evaluation information. These forms were applied in actual university lectures. We designed the system as a Web application to support both lecture forms. Figure 1 shows the perspectives of the students and teachers. Evaluation input from student handheld devices such as smartphones is enabled during a lecture through the system interface.

Figure 2 shows the system configuration. The system includes a UI for students and a UI for teachers. Individual students can input their lecture evaluation through the UI and a teacher uses the UI to check the lecture video as well as student feedback from the lecture evaluation.

Table 1. Evaluation forms

Type	FTF	After FTF	VOD	After VOD	Lecture form
A	Sequential	Overall			FTF using mobile terminals [1]
B	Sequential			Overall	Blended Learning(FTF+VOD)[2]
C		Overall	Sequential		
D			Sequential	Overall	e-Learning

FTF : face-to-face, VOD : video on demand

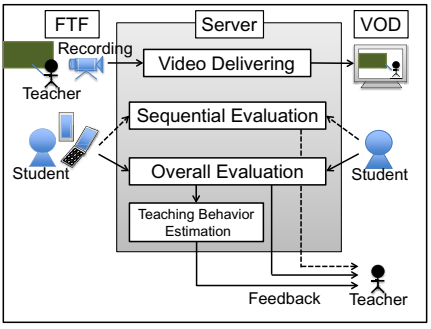

Fig. 1. Overview of evaluation forms

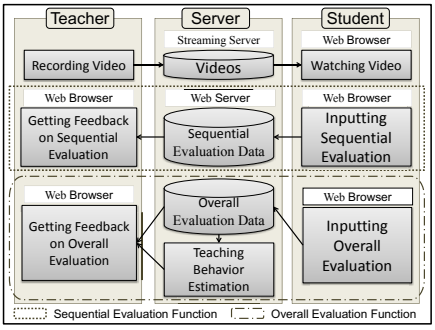

Fig. 2. System Configuration

3.1 Design of the Sequential Evaluation Function

One condition of sequential evaluation is that ratings change as the lecture progresses. A preliminary investigation was carried out to clarify what questionnaire items are suitable for sequential evaluation. We acquired one set of evaluation data during a certain lecture as it progressed and another set at the end of the lecture using questionnaire items from a general student survey. The results showed that six items were

available for sequential evaluation: “clarity of the explanation,” “degree of understanding of the lecture contents,” “degree of interest,” “teaching attitude,” “ease of hearing,” and “ease of seeing writing on the blackboard.”

It is necessary to design an input UI which, as much as possible, does not interfere with a lecture. Therefore, the system requires choices that students can choose intuitively. We carried out a preliminary survey to determine the choices, starting first with a prototype, then having students input their evaluation for the six questions while watching a lecture video. As a result, there were many opinions regarding whether a judgment was easy to immediately convey through the choices. For example, choices such as “I can understand” and “I’m not interested” were easy for students to choose. As a result, we decided to use these types of choices, and we developed a system experimentally and performed a preliminary experiment to have multiple students evaluate the six items. As a result of this preliminary experiment, we confirmed the possibility that a teacher could grasp areas for improvement in a lecture by paying attention to changes in the students’ evaluation. Thus, it is important that the teacher notices changes in the sequential evaluation information. To meet these requirements, we made a function to graph the time series change in the ratings; we call this function the time series graph function.

3.2 Design of the Overall Evaluation Function

We designed the feedback function of the overall evaluation data. The score of the overall evaluation made by a student does not highlight what the teacher should be paying more attention to. For example, it takes effort to pinpoint a problem with the blackboard demonstration when there are many evaluations saying it is “hard to understand a blackboard demonstration.” In contrast, the teacher can pay attention to the action of underlining important information more if the data states “the writing on the blackboard does not include enough emphasis through underlining.” For a student who has no professional teaching experience, it is difficult to directly evaluate teaching behavior. However, the overall evaluation from the student is affected by the teaching behavior that the teacher chose. This shows the possibility that the teacher’s teaching behavior can be estimated using an overall evaluation. Thus, in Section 4, we analyze the correlation between the overall evaluation and teaching behavior. We build a model to estimate teaching behavior from an overall evaluation based on the results of the analysis. Using this model, we suggest a function to feedback the teaching behavior evaluation information to the teacher. We call this function the teaching behavior estimation function.

4 Correlation Analysis between Teaching Behavior and Overall Evaluation

It is thought that the overall evaluation can be divided into “content evaluation” and “teaching behavior evaluation.” Thus, we expect a correlation between the overall evaluation and teaching behavior. We define teaching behavior in order to count it

and clarify the correlation between it and the overall evaluation. Next, we check this expectation and build a correlation model of overall evaluations and teaching behaviors. The following questions are examined:

4. Does teaching behavior affect the overall evaluation?
5. Can the model estimate individual teaching behaviors from among the many teaching behaviors displayed?

4.1 Method

We compared two lectures with similar content, one given in 2010 and the other in 2011, in order to determine whether specific teaching behaviors influence the overall evaluation. Overall evaluations were carried out during the information mathematics (IM) lecture given at the authors' universities. Two sets of data were collected: lectures 2–13 in 2010 and lectures 4–13 in 2011. Around 80 students attended each of the lectures. Twenty questionnaire items, such as "degree of content understanding" and "degree of interest," were chosen from an existing general student evaluation. For each item, the students answered using a 5-point Likert scale. The students were given approximately 5 minutes in each lecture to fill out the survey. The lecture was recorded with a video camera.

We first watched the lecture videos to identify the teaching behaviors used, then extracted two lectures of similar content that had a difference in rating level for a particular questionnaire item. We considered we would be able to identify how a specific teaching behavior affects the overall evaluation by comparing the use of the behavior in the two lectures. We performed a t-test to more specifically evaluate any questionnaire item identified. If there was at least one item with a difference in level of significance greater than 0.05, one of the authors watched the two lectures in detail and compared the evaluation results to the actual occurrence of the teaching behavior. In this way we determined which specific teaching behaviors to include in the analysis. We defined the specific teaching behaviors as those that could be counted from the video analysis and then counted each in turn. For example, we saw that changing the color of a phrase or sentence ("Coloring") and underlining a phrase or a sentence ("Underlining") frequently occurred in comparison with other teaching behaviors. Here "Coloring" denotes "the number of times that a color was changed in a word or sentence" and "Underlining" denotes "the number of times a word or sentence was underlined." Based on these definitions, one of the authors watched the paired lecture videos and counted the occurrences of coloring and underlining in each lecture. We confirmed whether a difference emerged in the overall evaluation results between the similar lectures by noting the difference in the number of times each teaching behavior occurred.

4.2 Results and Discussion

In comparing similar lectures, the score for a specific questionnaire item tended to increase if the teacher's activity at the blackboard increased. A specific example of this is the teacher at the blackboard "emphasizing a phrase or sentence" by coloring or

underlining, as defined above. Once the teaching behaviors of "Coloring" and "Underlining" were determined, we confirmed whether a difference emerges for the overall evaluation according to a difference in the number of times each teaching behavior occurred. The tenth lecture from the 2011 dataset was compared with the ninth lecture from the 2010 dataset as a large difference in coloring and underlining behavior was observed for lectures on the same topic. The frequency of both "Coloring" and "Underlining" was 0 in the ninth lecture in 2010, while the frequency of "Coloring" had increased to 3 and the frequency of "Underlining" had increased to 18 in the tenth lecture in 2011. T-test results showed the items that varied in the evaluations were "ease of notetaking" ($t=2.79$, $p\leq0.01$) and "teacher's ability to explain an abstract topic" ($t=2.53$, $p\leq0.01$). The lecture in 2011 clearly had a higher evaluation than for the similar lecture in 2010, and similarly there was a large difference in the number of times each teaching behavior was observed between the two lectures. Thus, these results indicate that these two different teaching behaviors affected the overall evaluation.

Correlation analysis was performed for all the lectures mentioned above, and models to estimate occurrences of "Coloring" and "Underlining" were developed. The questionnaire items indicating a slight correlation with "Coloring" were "Did the teacher explain abstract concepts plainly?" ($r = 0.60$), "Was the explanation simple?"($r = 0.49$), "Did the example aid your understanding?"($r = 0.46$), and "Was the quiz difficult?"($r = -0.47$). The questionnaire items indicating a slight correlation with "underlining" were "Did you feel the teacher's was motived in teaching?"($r = 0.53$) and "Did the teacher give an explanation that was easy to note down?"($r = 0.41$). The items that correlated to "Coloring" and "Underlining" were different, and therefore "Coloring" and "Underlining" can be estimated individually. A function to estimate "Coloring" and "Underlining" was then generated with those questionnaire items.

5 System Development

5.1 Time Series Graph Function

Based on the system design, a sequential evaluation input function and time series graph function were developed. Figure 3 shows the time series graph. This graph corresponds to conditions selected in the upper part. The values are plotted at the points the students evaluated, then connected by a straight line. The function can overlay evaluation results from multiple items or multiple students, as well as display the average rating.

If there were multiple parts of a lecture video for a teacher to watch, it was difficult to determine which parts should take precedence based only on the time series graph. As a supporting function to determine precedence, we used thumbnails of the lecture video. The thumbnail was connected to the evaluation point in the time series graph. The teacher could quickly confirm the contents of a specific part based on the content of the blackboard. In addition, the teacher could focus on improving areas only in the range of the video that corresponded to the thumbnail.

5.2 Teaching Behavior Estimation Function

Figure 4 shows the UI of the teaching behavior estimation function. Teachers can see messages on the upper portion and a scatter diagram on the lower portion. In the scatter diagram, the x-axis represents the occurrence of a teaching behavior and the y-axis represents the mean rating of the question. Actual data is drawn in the scatter diagram. The teacher can compare current data with past data. The message is changed based on the occurrences of a given teaching behavior measured in the past. The range between 0 and the maximum occurrence of a given teaching behavior is divided into 3 parts: "not performed very often," "performed a little," and "performed," depending on the estimated number of occurrences of the behavior.

"Degree of understanding of the content" and "Satisfaction with the lecture" are important standards as to whether or not a teacher should improve a teaching behavior. Therefore, to support the teaching behavior estimation function, we developed a function to show a rating ratio as a 100% accumulated stick graph for each question item.

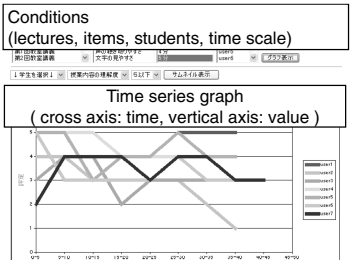

Fig. 3. A Time series graph

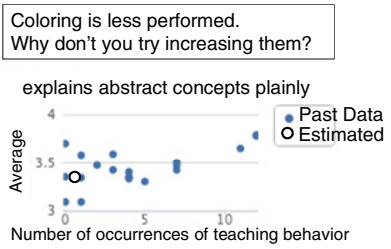

Fig. 4. User interface of teaching behavior estimation function

6 Experiments

Although it is true that teachers can grasp the points of improvement of a lecture more efficiently by combining the time series graph function and teaching behavior estimation function, we aim to clarify the usefulness of each function in this research. Therefore, individual evaluation experiments were carried out for the time series graph function and for the teaching behavior estimation function.

This system was designed for use with multiple lecture forms. However, in order to evaluate the functions, we imposed three restrictions in the experimental conditions so that students could perform as many evaluations as possible:

1. Students input sequential evaluation in an e-learning environment.
2. The system specifies the evaluation timing while the students evaluate all six items simultaneously.
3. The video stops while the student is evaluating.

Overall evaluations experiments were carried out in face-to-face (FTF) lectures in order to secure participants.

6.1 Evaluation of the Time Series Graph Function

Evaluation of the function is based on a teacher's subjective evaluations. There are two teachers: Teacher A and Teacher B. We recorded in advance the multiple lectures they were in charge of teaching. We chose one video of a difficult lecture topic from Teacher A (Lecture A), and the corresponding lecture video from Teacher B (Lecture B). Lecture A is 90 minutes long and Lecture B is 45 minutes long, and the evaluation intervals of Lecture A and Lecture B are 10 minutes and 5 minutes, respectively. Fourteen students in all participated in this experiment: seven students evaluated Lecture A and the other seven students evaluated Lecture B. All students were college seniors at the authors' universities.

Teachers A and B then checked the data using the time series graph function. We first gave them directions on how to use the function, then had them use the system individually. They checked whether their lecture had any points for improvement. Finally, we distributed a questionnaire to them to evaluate the function. The questions on the questionnaire were as follows:

Qs1: Do you have any requests regarding the evaluation timing?
Qs2: Do you want to actually use this system?
Qs3: Do you think that a time series graph function is useful?
Qs4: Do you think that a thumbnail function is useful?
Qs5: Can you grasp the lecture points for improvement using this system?

The answers for Qs2 through Qs5 were given using a 5-point Likert scale. Teacher A answered "4" for all questions. Teacher B answered "4" for Qs5 and "5" for all other questions. Both teachers submitted positive evaluations. In explaining their responses to Qs3 and Qs4, they expressed that the system was useful because it would be difficult to notice a bad portion in the video and it would also take time if watch the video from the beginning. These results show the effectiveness of the time series graph to support teachers in pinpointing points for improvement from a lecture video. As for Qs1, their opinion was that it would be even more useful if the timing could be specified as arbitrary time in the lecture video. On Qs5 they said that it takes a little more time to find a clear point for improvement, although it was possible to narrow it down to some extent. Given these comments, a tool for more effective teaching behavior support should be developed in the future. However, as this experiment was conducted under the limited conditions, in future research it is necessary to run the experiment in a more realistic situation with fewer conditions imposed.

6.2 Evaluation of the Teaching Behavior Estimation Function

Overall evaluations of the teaching behavior estimation function were carried out during IM lectures at the authors' universities in the 2012 fiscal year. The evaluation considered 10 lectures. We used 20 items for the construction of the model identical to those described in Section 4. There were an average of 74.1 participants (SD = 5.7), all in either the second or third grade. All the lectures were recorded. Teachers watched all 10 lecture videos, using the teaching behavior estimation function. The subjects were five teachers, consisting of the lead teacher (R) and four coworkers (P1, P2, A1,

A2). R is an associate professor with 10 years of work experience. P1 and P2 are professors with 27 and 31 years of experience. A1 and A2 are assistant instructors with 1 and 4 years of work experience. Each teacher evaluated the teacher behavior estimation function after watching the video. The teachers' subjective evaluations were obtained with a questionnaire. The reason we targeted fellow teachers is to investigate whether the evaluation changed depending on the teacher's background.

Table 2. Results of the teachers' evaluation of the teaching method

#	Question	R	P1	P2	A1	A2
Qo1	The teaching method affects students' understanding of lecture contents.	5	4	4	4	5
Qo2	Students can judge whether a quality of teaching is good or bad.	4	4	4	2	2
Qo3	Student evaluations are useful for improving lectures.	4	4	4	4	5
Qo4	The lesson evaluation acquired after each lecture is effective for lesson improvement.	4	2	3	4	5
Qo5	It is useful for a teaching point for improvement to be estimated by the students' lecture evaluation.	5	3	5	5	5

Table 3. Evaluation of the teacher behavior estimation function by teachers

#	Question	P1	P2	Avg.	R	A1	A2	Avg.
Qo6	Lecture videos with a message and the scatter diagram about "Underlining" help to pinpoint the areas for improvement in the lecture.	2	2	2	3	3	4	3.5
Qo7	Lecture videos with a message and the scatter diagram about "Coloring" help to pinpoint the areas for improvement in the lecture.	2	2	2	3	3	4	3.5
Qo8	The message about "Underlining" served as a reference in examining the points for improvement in the lecture.	2	2	2	3	4	4	4
Qo9	The message about "Coloring" served as a reference in examining the points for improvement in the lecture.	2	2	2	3	4	4	4
Qo10	The scatter diagram about "Underlining" served as a reference in examining the points for improvement in the lecture.	2	2	2	2	3	4	3.5
Qo11	The scatter diagram about "Coloring" served as a reference in examining the points for improvement in the lecture.	2	2	2	2	3	4	3.5
Qo12	The bar graph, message, and scatter diagram about "Underlining" are useful in examining the points for improvement in the lecture.	2	2	2	4	4	5	4.5
Qo13	The bar graph, message, and scatter diagram about "Coloring" are useful in examining the points for improvement in the lecture.	2	2	2	4	4	5	4.5
Qo14	The message about "Underlining" was appropriate.	3	2	2.5	4	3	5	4
Qo15	The message about "Coloring" was appropriate.	3	3	3	4	3	5	4

Table 2 shows the evaluations of using teaching behaviors for lecture improvements submitted by the teachers. Nearly all the teachers answered positively to Qo1 and Qo5. Teachers A1 and A2 answered Qo2 negatively, explaining that although students can perform subjective evaluations, it is difficult for them to evaluate teaching behavior. The Qo3 and Qo4 results show that lecture evaluations alone are not effective for lecture improvement. These results suggest that the teachers support the use of the teaching behavior estimation function.

Table 3 shows the results of the teachers' evaluation of the teaching behavior estimation function. In comparing all items, R has a tendency to rate higher than P, and A has a tendency to rate higher than R. Teacher P1 provided this reason for a negative evaluation: "Although it is advisable, it is not essential." On the contrary, R said "it is useful to know a tendency for a teaching behavior because this leads to enhanced

awareness." Moreover, A1 had the same opinion as R. According to these results, although the teachers with long work experience regard the information as important, the teachers with comparatively short work experience tend to view the information as constructive criticism that can be used to improve their lectures. Therefore, the teaching behavior estimation function may be useful, especially for young teachers.

We interviewed R about his intention when underlining and changing the color of a character. He replied that he uses underline unconsciously, but is aware of consciously changing the character color to distinguish a concept. This shows that the correlation between "Coloring" and "explains an abstract concept plainly" is a valid relation.

7 Conclusions and Future Work

We have proposed a function that provides sequential evaluation and overall evaluation information to the teacher as feedback. A system with such functions was developed and an evaluation experiment was conducted on it. We showed through subjective evaluations from teachers that the sequential evaluation and overall evaluation information was helpful. The advantage of the time series graph function was that it was able to pinpoint portions of the lecture videos that should be reviewed. The teaching behavior estimation function was useful in making teachers more aware of their teaching behaviors. A teaching behavior estimation function could possibly be useful, especially for teachers with little work experience.

The evaluation of the time series graph function and the evaluation of the teaching behavior estimation function were carried out separately. In the future we will examine how to combine these functions, increase the types of teaching behavior examined, and build a higher-precision model.

Acknowledgements. This research was partially supported by the Japanese Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research, 24300291, 2012–2015.

References

1. Nagaoka, K.: A response analyzer system utilizing mobile phones. In: Proc. of the IASTED International Conference Web-Based Education, pp. 579–584 (2005)
2. Nikolaidou, M., Sofianopoulou, C., Giannopoulos, I.: Assessing the Contribution of Lecture Video Service in the Hybrid Learning Ecosystem of Harokopio University of Athens. In: 2010 Second International Conference on Mobile, Hybrid, and On-Line Learning, pp. 141–146 (2010)
3. Stalmeijer, R.E., Dolmans, D.H.J.M., Wolfhagen, I.H.A.P., Peters, W.G.: Lieve van Coppenolle, Scherpbier, A. J. J. A.: Combined student ratings and self-assessment provide useful feedback for clinical teachers. Adv. Health Sci. Educ. 15, 315–328 (2010)
4. Hanakawa, N., Obana, M.: Lecture Improvement based on Twitter Logs and Lecture Video using p-HInT. In: Proc. of the 18th International Conference on Computers in Education, pp. 328–335 (2010)