Assessment of Educational Effectiveness in Real–Time Distance Lectures over IP Networks

I. INTRODUCTION

Owing to the spread of high–speed networks and high performance terminals, many applications which rely on audio–video transmission over the Internet become popular. Distance learning applications fall into this category.

The distance learning applications can be classified into two types: on–demand and real–time [1]. The on–demand type is a mode of online delivery where students access course materials (e.g., recorded–video) on their own schedule. The students are not required to be together at the same time and can repeatedly watch the lecture. The real–time type is online learning which all students “present” themselves at the same time in different places. This type commonly makes use of TV conferencing system; a teacher and all students have to participate in the lecture at its scheduled time. The teacher can watch students’ head nodding and expressions. Then, he/she can speak at an appropriate speed for the students; this realizes a kind of interactivity. In this paper, we deal with the real–time distance lecture.

The enhancement of educational effectiveness is very important in the distance lectures. The educational effectiveness is affected by the audio–video quality. However, the current Internet can provide the best effort service only. Consequently, QoS (Quality of Service) of audio–video transmission cannot be guaranteed over the Internet. Therefore, QoS control is required.

In audio–video transmission over IP networks, its temporal structure can be easily disturbed by delay jitter of packets. The disturbance decreases fidelity of the audio–video stream. The impairment of the fidelity can be remedied by a playout buffer [2] in the receiver; packets which arrive at the receiver are stored in its buffer so that the delay jitter can be absorbed. However, the playout buffer increases latency because of the buffering time.

References [3] and [4] show that the effect of playout buffering time on QoE (Quality of Experience), which is quality perceived by the user [5], in interactive audiovisual communications over IP networks. Reference [3] examines the effect of playout buffering control on QoE over best–effort IP networks. Reference [4] treats audiovisual communications over bandwidth guaranteed IP networks. The studies confirm that we can achieve high QoE by choosing an appropriate playout buffering time. However, the studies do not target educational applications and then do not presume the users which have special roles, i.e., a teacher, students and teaching assistants. Moreover, they do not assess the effectiveness from an educational point of view.

The spatial quality of video also deteriorates owing to video encoding and packet loss in distance lectures over IP networks; they influence students’ educational effectiveness. In [6], Fujiki et al. investigate the influence of video encoding bit rates and network load traffic on the ratio of the correct answers in the test and subjective assessment. Moreover, they also assess the influence of audio encoding bit rates and the network load traffic. However, the study supposes an on–demand distance lecture; that is, they use recorded–video and recorded–audio in the experiment. Therefore, the influence of interactivity between a teacher and students on educational effectiveness is not clarified. In addition, they do not investigate the effect of playout buffering time on educational effectiveness.

In this paper, we deal with distance lectures over IP networks and assess educational effectiveness for various values of the video encoding bit rate and playout buffering time by experiment. We assume a lecture with a blackboard. The teacher gives a lecture while seeing students who stay in another room; this implies that low interactivity exists between the teacher and students. For simplicity of discussion, however, we do not treat question–and–answer between the teacher and students (i.e., high interactivity) in the experiment since this is a first step of the study.

The educational effectiveness is evaluated in terms of satisfaction and understanding. In order to investigate the satisfaction in a multidimensional way, we utilize the SD (Semantic Differential) method [7], which is one of the psychometric methods. Moreover, we also assess student’s overall satisfaction for the distance lecture. In order to clarify relationship between each factor (e.g., video quality and students’ motivation) and the overall satisfaction, we calculate the coefficient of correlation between each result of the multidimensional assessment and the overall satisfaction.

As the metric of the understanding, we adopt the learning quotient (growth rate), which is defined as the ratio of the correct answer in the pre–test to that in the post–test. We also clarify relationship between each result of the multidimensional assessment and the learning quotient.

The rest of the paper is organized as follows. Section II introduces a method of assessing educational effectiveness. Section III presents our experiment. We show our experimental results and consideration in Section IV. Section V concludes the paper.

II. EDUCATIONAL EFFECTIVENESS ASSESSMENT

A. SD method

The SD method was proposed by Osgood as a method of measuring meaning. This method can assess an object for evaluation from many points of view with many pairs of polar terms. A pair of polar terms consists of one adjective and its opposite one; e.g., quiet and noisy.

In the SD method, how to select pairs of polar terms used for assessment is important. For each selected pair of polar terms, a subjective score of an object for evaluation is measured by the rating–scale method [8]. The rating–scale method is also used to measure MOS (MOS: Mean Opinion Score), which is widely utilized for assessment of a single media. We refer to an object for evaluation as a stimulus.
In the rating–scale method, subjects classify each stimulus into one of a certain number of categories. Each category has a predefined number. For example, "good" is assigned 5, "fair" 4, "poor" 3, "bad" 2 and "horrible" 1. However, the numbers assigned to the categories only have a greater–than–less-than relation between them; that is, the assigned number is nothing but an ordinal scale. When we assess the subjectivity quantitatively, it is desirable to use at least an interval scale.

In order to obtain an interval scale from the result of the rating–scale method, we first measure the frequency of each category with which the stimulus is placed in the category. With the law of categorical judgment [8], we can translate the frequency obtained by the rating–scale method into an interval scale. Since the law of categorical judgment is a suite of assumptions, we must test goodness of fit between the obtained interval scale and the post-test. Mosteller [9] proposed a method of testing the goodness of fit for a scale calculated with Thurstone’s law of comparative judgment [8], which is one of psychometric methods. The method can be applied to a scale obtained by the law of categorical judgment. This paper uses Mosteller’s method to test the goodness of fit.

**B. Students’ understanding assessment**

In this paper, we assess the subjects’ understanding by a pre-test and a post-test. We ask all students the same questions in the pre-test and the post-test. The ratio of the correct answer in the pre-test is denoted as \( a \) \( \% \), while that in the post-test as \( b \) \( \% \). We obtain the learning quotient (LQ) by the following expression.

\[
LQ = \frac{b - a}{100 - a} \times 100
\]

In this paper, we utilize the recall–test [10] in the pre-test and the post-test. In the recall–test, the student fills in a white paper or blanks on the basis of his/her memory. The recall–test is divided into two groups: simple–recall–test and completion–test. The simple–recall–test is a method by which the student fills in a white paper, while the completion–test instructs the student to fill in blanks in sentences, expressions and figures. We use the completion–test because an outline is shown in an answer sheet; the students can answer questions even by watching a lecture once.

### III. EXPERIMENT

This section explains the experimental system and assessment method. In the experiment, we used two classrooms: a lecture classroom and a distance classroom. In the lecture classroom, a teacher gave lectures. In the distance classroom, students took the lectures.

**A. Experimental system**

Figure 1 shows the configuration of the experimental system. It consists of two routers and six PCs, which are used as load senders, load receivers and two terminals. Routers 1 and 2 are RiverStone’s RS3000. The links in the system are all Ethernet channels. The link speed between a PC and a router is 100 Mb/s, while that between the two routers is 10 Mb/s. The link between the two routers is therefore a bottleneck. Each terminal is equipped with a real–time video encoding board made by the DSP Research, Inc. In the board, the difference between the target bit rate and the actual average encoding bit rate is less than 10 \%

In the lecture classroom, the video was projected on the screen; we put the screen for the teacher so that he can see the students. We distributed a white paper for notes to each student. Table I shows the specification of the transmitted audio and video. In Table I, MU stands for the media unit, which is the information unit for media synchronization at the application level. In this paper, we define a video frame as a video MU and a constant number of audio samples as an audio MU.

Load Senders 1 and 2 transmitted load traffic to Load Receivers 1 and 2, respectively. Each load sender generated fixed–size UDP datagrams of 1472 bytes each in its payload at exponentially distributed intervals and sent them to the corresponding load receiver. We set the average amount of load traffic to 3.0 Mb/s in each load sender.

We utilized a simple playout buffering control scheme. The buffering time was set to 100, 200, 400 and 800 ms. Terminals 1 and 2 used the same buffering time.

**B. Lecture content**

The lecture content was “Electronic measurements and controls,” which is taken by students of electrical and electronic engineering in technical high schools. The reason why we used the content is as follows. In the experiment, the subjects were students who major in computer science. Therefore, to reduce the influence of the precedency knowledge, we employed a lecture content which is not in their major; we used a textbook entitled Electronic Measurement Control [11]. One lecture time was set to 10–15 minutes in consideration of student’s burden.

We had 12 lectures because of three video encoding bit rates and four values of the buffering time. Before the lectures, the teacher gave the students a sample lecture for training of the assessment. In the sample lecture, we set the video encoding bit rate to 3 Mb/s and the playout buffering time to 400 ms. Thus, the students took the 13 lectures. The 13 lectures are related to each other in the content. In particular, students’ understanding in a lecture is affected by the previous lectures. Therefore, in the experiment, the students were divided into two groups. The order of 12 combinations of encoding bit rates and buffering time differs for each group.

![Fig. 1. System configuration.](image-url)
C. Selection of pairs of polar terms

We selected pairs of polar terms of Japanese adjectives for the SD method. In order to investigate how students recognize audio quality, video quality, lecture content and student’s motivation separately, we explicitly specified the subject (i.e., audio, video) by adding the type to each adjective. When we could not find any appropriate adjective, we adopted a verb instead.

We collected 60 pairs of polar terms and classified them into 11 categories: A (Audio: 9 pairs), B (Video: 10 pairs), C (Audio and Video synchronization: 1 pair), D (Interactivity: 2 pairs), E (Presence: 5 pairs), F (Tiredness: 3 pairs), G (Tension: 2 pairs), H (System: 10 pairs), I (Student’s motivation: 9 pairs), J (Lecture content: 9 pairs) and K (Subjective learning effect: 1 pair). There is a pair which is related to two categories.

Note that this experiment was performed in Japanese. This paper has translated the used Japanese pairs of polar terms into English. Therefore, the meanings of adjectives or verbs written in English here slightly differ from those of Japanese ones. For convenience, we have assigned an identification alphabet and number to each pair of polar terms as shown in Tables II and III.

D. Overall quality assessment

The students assessed the overall satisfaction of the distance lecture by the rating–scale method. In the rating–scale method, we used ACR(Absolute Category Rating). We defined five categories: “excellent” is assigned integer 5, “good” 4, “fair” 3, “poor” 2, and “bad” 1. We applied the law of categorical judgment to the obtained result.

E. Multidimensional assessment with SD method

The students assessed the 12 lectures with the rating–scale method for each pair of polar terms. In order to express the degree of implication of the adjectives or verbs, we used two kinds of adverb and one adjective: very, slightly and neutral. As a result, we had five terms for each adjective or verb: “one adjective or verb with very”, “the one with slightly”, “neutral”, “the opposite one with slightly” and “the one with very”. For convenience, we refer to the terms as categories 5, 4, 3, 2 and 1. We assume that the term means higher quality as the category number becomes larger. The students are male and female in their twenties. The number of students is 16.

F. Experimental procedure

The experiment was conducted as follows.

1) The students take a pre–test without time constraint. We forbid the students to read and write notes.
2) The students take a lecture. We permit students to take notes of a lecture. However, we forbid the students to read and write notes.
3) The students take a post–test without time constraint. We forbid the students to read and write notes.
4) The students assess subjective quality without time constraint. We forbid the students to read and write notes.
5) We collected 60 pairs of polar terms and classified them into 11 categories: A (Audio: 9 pairs), B (Video: 10 pairs), C (Audio and Video synchronization: 1 pair), D (Interactivity: 2 pairs), E (Presence: 5 pairs), F (Tiredness: 3 pairs), G (Tension: 2 pairs), H (System: 10 pairs), I (Student’s motivation: 9 pairs), J (Lecture content: 9 pairs) and K (Subjective learning effect: 1 pair). There is a pair which is related to two categories.

We removed three values which give large errors of Mosteller’s test and then found that the test with a significance level of 0.05 cannot reject the hypothesis. Therefore, we consider the interval scale as the psychological scale.

Since we can select an arbitrary origin in the interval scale, we set the lower boundary of category 2 to the origin. Under this condition, we calculated the lower boundaries of the categories and got 2.086 for Category 5, 2.018 for Category 4, 0.987 for Category 3, and 0 for Category 2.

Figure 2 plots the overall psychological scale value versus the playout buffering time for the three video encoding bit rates; it should be noted that the results of stimuli removed by Mosteller’s test are not shown. We see in the figure that the overall psychological scale values increase as the video encoding bit rate increases for all the buffering time considered here. As the video encoding bit rate increases, the spatial quality of video improves. Thus, we confirm that the spatial quality of video affects students’ overall satisfaction.

In Fig. 2, we also find that the psychological scale value increases as the buffering time increases at each video encoding bit rate. This is because the video quality improves with increase of buffering time. In this experiment, we have not treat question–and–answer between the teacher and the students; that is, there exists only low interactivity. In such a situation, they hardly notice the degradation of interactivity due to the latency.

B. Psychological scale for each pair of polar terms

In order to obtain the interval scale for each pair of the polar terms, we removed some values which give large errors of Mosteller’s test with a significance level of 0.05.

Next, we compute the correlation coefficients between the overall satisfaction and each pair of adjective for 60 ones. We apply t–authorization by significance level 0.01 to the result. Then, we obtain 20 adjective pairs. We display the 20 adjective pairs and the correlation coefficients in Table II.

In Table II, we find three pairs for A (Audio), seven pairs for B (Video), a pair for F (Tiredness), five pairs for H (System), a pair for I (Student’s motivation), and three pairs for J (Lecture content).

Especially, among the ten pairs for B (Video), seven pairs highly correlates with the overall satisfaction. Moreover, we find B-8 “Writing and figures are easy to see–Writing and figures are not easy to see” has the highest correlation coefficient among the adjective pairs for video quality. Therefore, the clearness of characters and figures on the blackboard strongly affects the overall satisfaction.

C. Learning quotient assessment

Figure 3 indicates the learning quotient as a function of the buffering time for the three video encoding bit rates. We find in this figure that the learning quotient does not have a clear relationship to the video encoding bit rate and the buffering time as in the overall satisfaction. The reason is as follows. In the experiment, we set the average amount of UDP load traffic to 3.0 Mb/s, which corresponds to a lightly loaded condition. Thus, the MU loss ratio is almost 0%. The students merely miss the teacher’s talk and characters on the blackboard by packet loss. Therefore, the students’ understanding is scarcely
TABLE II
CORRELATION OF OVERALL SATISFACTION WITH PAIRS OF POLAR TERMS.

<table>
<thead>
<tr>
<th>pair of polar terms</th>
<th>correlation coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>J-6</td>
<td>The lecture calm down--The lecture is hurried</td>
</tr>
<tr>
<td>B-8</td>
<td>Writing and figures are easy to see--Writing and figures are not easy to see</td>
</tr>
<tr>
<td>B-5</td>
<td>Video is not blurred--Video is blurred</td>
</tr>
<tr>
<td>H-6</td>
<td>The system is kind to learn--The system is not kind to learn</td>
</tr>
<tr>
<td>H-9</td>
<td>The system is practical--The system is impractical</td>
</tr>
<tr>
<td>B-6</td>
<td>Video is clear--Video is not clear</td>
</tr>
<tr>
<td>H-10</td>
<td>I want to participate in a distance lecture--I do not want to participate in a distance lecture</td>
</tr>
<tr>
<td>H-2</td>
<td>The system does not make me feel stress--The system makes me feel stress</td>
</tr>
<tr>
<td>B-4</td>
<td>Video is natural--Video is artificial</td>
</tr>
<tr>
<td>J-7</td>
<td>The lecture is easy to understand--The lecture is not easy to understand</td>
</tr>
<tr>
<td>J-5</td>
<td>The lecture is simple--The lecture is complicated</td>
</tr>
<tr>
<td>B-3</td>
<td>Video is stable--Video is unstable</td>
</tr>
<tr>
<td>B-9</td>
<td>Teacher's gestures and behavior is easy to see--Teacher's gestures and behavior is not easy to see</td>
</tr>
<tr>
<td>F-1</td>
<td>My eye does not feel tired--My eyes feel tired</td>
</tr>
<tr>
<td>A-2</td>
<td>Audio does not distort--Audio distorts</td>
</tr>
<tr>
<td>H-7</td>
<td>The system is effective--The system is ineffective</td>
</tr>
<tr>
<td>A-1</td>
<td>Audio does not have noise--Audio has noise</td>
</tr>
<tr>
<td>A-4</td>
<td>Audio is clear--Audio is not clear</td>
</tr>
<tr>
<td>I-6</td>
<td>I feel pleasant with the lecture--I do not feel pleasant with the lecture</td>
</tr>
<tr>
<td>B-1</td>
<td>Video does not drop out--Video drops out</td>
</tr>
</tbody>
</table>

Fig. 3. Learning quotient.

influenced by the video encoding bit rate and the buffering time.

Next, we calculate the correlation coefficient between the learning quotient and the psychological scale for the 60 adjective pairs. In order to find adjective pairs which have strong correlations with the learning quotient, we apply t-authorization by significance level 0.05 to the result. Then, we obtain seven adjective pairs. We display the seven adjective pairs and the correlation coefficients in Table III.

In Table III, we find a pair for A (Audio), a pair for F (Tiredness), a pair for G (Tension), three pairs for I (Student’s motivation), and a pair for J (Lecture content). This observation implies that when the network is lightly loaded, the learning quotient has a stronger relationship to the adjective pairs for student’s motivation than those for media quality, e.g., audio and video.

V. CONCLUSION

In this paper, we investigated the effect of video encoding bit rates and buffering time on educational effectiveness in audiovisual distance lectures over IP networks by experiment. As a result, we observed the following. The effect of the video encoding bit rate on the overall satisfaction is large. The adjective pairs for the video quality have a strong correlation with the overall satisfaction. In particular, the clearness of characters and figures on the blackboard highly correlates with the overall satisfaction. If there is no question-–answer between the teacher and the students even when the teacher gives a lecture while seeing students, i.e., only low interactivity exists, the overall satisfaction is not affected by latency so much.

The learning quotient has a strong relationship with the adjective pairs for student’s motivation and interest.

As future work, we will treat distance lectures with question-–answer and cooperative learning (e.g., group learning); they require high interactivity.

ACKNOWLEDGMENT

This work was supported by the Grant–In–Aid for Scientific Research of Japan Society for the Promotion of Science under Grant 21560183 and The Telecommunications Advancement Foundation.

REFERENCES