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

Koichiro Noda, Toshiro Nunome and Shuji Tasaka
Graduate School of Engineering,
Nagoya Institute of Technology
Nagoya 466–8555, Japan

Kazuyoshi Fukaya
School of Education,
Sugiyama Jogakuen University
Nagoya 464-8662, Japan

***Abstract*—This paper deals with audiovisual distance lectures over IP networks and assesses educational effectiveness for various values of the video encoding bit rate and playout buffering time by experiment. The educational effectiveness is evaluated in terms of satisfaction and understanding. We investigate the satisfaction in a multidimensional way by the SD method with many adjective pairs. The understanding is expressed by 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. As a result, we find that the satisfaction closely correlates with adjective pairs for picture quality. In addition, we notice that the learning quotient has a strong relationship to adjective pairs for student's motivation and interest.**

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(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, "excellent" is assigned 5, "good" 4, "fair" 3, "poor" 2 and "bad" 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. We refer to the interval scale as the *psychological 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 measurement result. 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 students' 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 100Mb/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 teacher gave a lecture by using a blackboard. Terminal 1 transmitted a pair of audio and video streams to Terminal 2 with UDP. In the distance classroom, the video was projected on the screen. The audio was output by a speaker which is equipped with the classroom.

Similarly, Terminal 2 transmitted the video stream to Terminal 1 with UDP. Note that Terminal 2 did not transmit the audio stream because we have not treated question–and–answer between the teacher and the students in this experiment. 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.

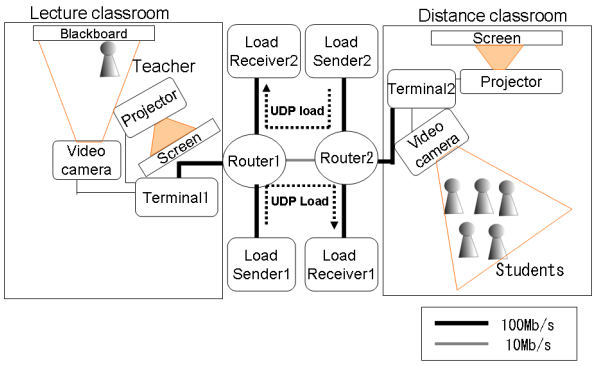

Fig. 1. System configuration.

TABLE I
AUDIO AND VIDEO SPECIFICATIONS.

	audio	video
coding method	ITU–T G.711 μ–law	H.264
average bit rate [kb/s]	64	1000, 3000 6000
image size [pixel]	–	320 × 240
picture pattern	–	I
average MU interval [ms]	40	40
average MU rate [MU/s]	25	25

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 precedence 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.

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 evaluation papers.
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) Return to 1), if the next lecture exists.

We directed the students to assess subjective quality after post–test. It took about 330 minutes for each group to finish all the assessment.

IV. Experimental Results

A. Overall quality assessment

We assessed the overall perceptual quality of the 12 lectures by the rating–scale method. By applying the law of categorical judgment to the result, we calculated the interval scale which indicates the overall satisfaction. As a result of Mosteller's test [9], the hypothesis that the observed value equals the calculated one was rejected with a significance level of 0.05. 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 2.986 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

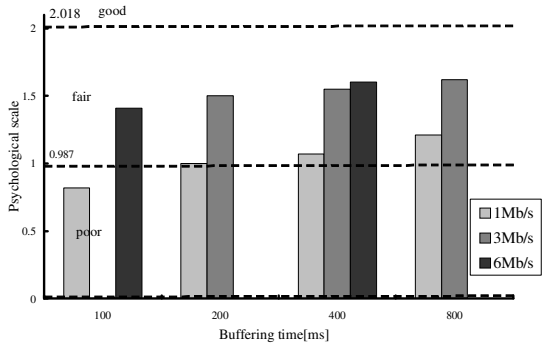

Fig. 2. Overall psychological scale.

TABLE II
CORRELATION OF OVERALL SATISFACTION WITH PAIRS OF POLAR TERMS.

pair of polar terms		correlation coefficient
J-6	The lecture calm down–The lecture is hurried	0.917
B-8	Writing and figures are easy to see– Writing and figures are not easy to see	0.912
B-5	Video is not blurred–Video is blurred	0.874
H-6	The system is kind to learn– The system is not kind to learn	0.870
H-9	The system is practical–The system is impractical	0.867
B-6	Video is clear–Video is not clear	0.848
H-10	I want to participate in a distance lecture– I do not want to participate in a distance lecture	0.840
H-2	The system does not make me feel stress– The system makes me feel stress	0.813
B-4	Video is natural–Video is artificial	0.800
J-7	The lecture is easy to understand– The lecture is not easy to understand	0.767
J-5	The lecture is simple–The lecture is complicated	0.757
B-3	Video is stable–Video is unstable	0.742
B-9	Teacher's gestures and behavior is easy to see– Teacher's gestures and behavior is not easy to see	0.712
F-1	My eyes do not feel tired–My eyes feel tired	0.708
A-2	Audio does not distort–Audio distorts	0.696
H-7	The system is effective–The system is ineffective	0.676
A-1	Audio does not have noise–Audio has noise	0.650
A-4	Audio is clear–Audio is not clear	0.634
I-6	I feel pleasant with the lecture– I do not feel pleasant with the lecture	0.629
B-1	Video does not drop out–Video drops out	0.624

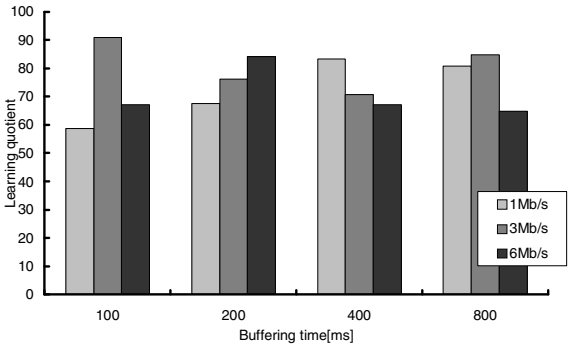

Fig. 3. Learning quotient.

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

TABLE III
CORRELATION OF LEARNING QUOTIENT WITH PAIRS OF POLAR TERMS.

pair of polar terms		correlation coefficient
F-3 G-1	I am not getting sleepy–I am getting sleepy	0.636
I-8	I am fascinated by the lecture– I am not fascinated by the lecture	0.580
J-5	The lecture is simple–The lecture is complicated	-0.562
A-8	Audio is rich–Audio is poor	0.549
I-4	I enjoy learning in the lecture– I do not enjoy learning in the lecture	0.533
I-9	I am interested in the lecture– I am bored in the lecture	0.502
H-5	The system is reliable–The system is unreliable	0.500

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 not question–and–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–and–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 21360183 and The Telecommunications Advancement Foundation.

REFERENCES

[1] S. Hrastinski, "Asynchronous and synchronous e–learning," *Educause quarterly*, vol. 31, no. 4, Dec. 2008.

[2] K. Fujimoto, S. Ata and M. Murata, "Adaptive playout buffer algorithm for enhancing perceived quality of streaming applications," *Conf. Rec. IEEE GLOBECOM2002*, pp. 2463–2469, Nov. 2002.

[3] Y. Ito and S. Tasaka, "Tradeoff relationship between fidelity and latency in interactive audio–video applications over IP networks," *IEICE Trans. Commun.*, vol. E90–B, no. 5, pp.1112–1121, May 2007.

[4] N. Misaki and S. Tasaka, "QoE assessment of interactive audio–video communications over bandwidth guaranteed IP networks," (in Japanese) *IEICE Tech. Rep.*, CQ2008–7, pp.35–40, Apr. 2008.

[5] ITU–T Rec. G.100/P.10 Amendment 1, "Amendment 1:New Appendix 1 – Definition of Quality of Experience (QoE)," Jan. 2007.

[6] T. Fujiki, M. Murota and Y. Shimizu, "Evaluation as teaching media on video conference tool for low bit rate," (in Japanese) *IEICE Tech. Rep.*, ET2001-82, pp.115–122, Dec. 2001.

[7] C. E. Osgood, "The nature and measurement of meaning," *Psychological Bulletin*, vol. 49, no. 3, pp. 197–237, May 1952.

[8] J. P. Guilford, *Psychometric methods*, McGraw–Hill, N. Y., 1954.

[9] F. Mosteller, "Remarks on the method of paired comparisons: III. a test of significance for paired comparisons when equal standard deviations and equal correlations are assumed," *Psychometrika*, vol. 16, no. 2, pp.207–218, June 1951.

[10] F. Haist, A. P. Shimamura and L. R. Squire, "On the relationship between recall and recognition memory," *Journal of Experimental Psychology: Learning, Memory and Cognition*, vol. 18, no. 4, pp.691-702, July 1992.

[11] Y. Tsuzuki, Y. Ikeda, K. Kobayashi, T. Aoki and I. Izumi, "*Electronic Measurement Control*,"(in Japanese) CORONA Publishing Co., LTD., 2003.