Multimedia Systems 7: 486–499 (1999)

Multimedia Systems

Experimental evaluation of loss perception in continuous media*

Duminda Wijesekera[1], **Jaideep Srivastava**[1], **Anil Nerode**[2], **Mark Foresti**[3]

[1] Department of Computer Science, University of Minnesota, Minneapolis, MN 55455, USA (e-mail: {wijesek,srivasta}@cs.umn.edu, forestim@rl.af.mil)
[2] Institute for Intelligent Machines, Cornell University, Ithaca, NY 14853, USA
[3] Rome Laboratory, Griffis Air Force Base, Rome, NY 14853, USA

Abstract. Perception of multimedia quality, specified by quality-of-service (QoS) metrics, can be used by system designers to optimize customer satisfaction within resource bounds enforced by general-purpose computing platforms. Media losses, rate variations and transient synchronization losses have been suspected to affect human perception of multimedia quality. This paper presents metrics to measure such defects, and results of a series of user experiments that justify such speculations. Results of the study provide bounds on losses, rate variations and transient synchronization losses as a function of user satisfaction, in the form of Likert values. It is shown how these results can be used by algorithm designers of underlying multimedia systems.

Key words: Quality of service – User studies – Media losses – Metrics

1 Introduction

Multimedia systems, characterized by integrated computer-controlled generation, manipulation, presentation, storage and communication of independent discrete and continuous-media (CM) data [SGN96], have to compete for the same clientele that has already been accustomed to high standards set by radio and broadcast television. It is a challenge to provide services comparable to broadcast TV on general-purpose computing platforms, since the former is based on an architecture supported by dedicated channels. Fortunately, due to inherent limitations of human perception, some loss of quality can be tolerated. Hence, it is sufficient to provide multimedia services to be within such tolerable limits. The *goodness* of a presentation is called its quality of service (QoS) in the multimedia literature. While a number of mathematical measures of QoS have been proposed in the literature [WS96, Tow93], there is hardly any work on validating them through measurements of human perception. The need for performing such studies has been expressed in a number of papers, e.g., [SGN96, Geo96, SB96]. The current paper reports results of some experiments in measuring human tolerance to lossy media.

Two widely quoted papers on user studies of multimedia systems are [Ste96] and [AFKN94]. Based on an extensive study, [Ste96] concluded that audio-video lip-synchronization errors of 80 ms were undetectable, up to 120 ms were detectable but tolerated, and above 120 ms were irritating. For audio-pointer synchronization, the respective limits were 200 and 1000 ms. In [AFKN94], perceptual effects of different frame rates were investigated for audio-visual clips with high temporal, audio and video content, i.e., having high-speed action.

Both these experiments were carried out for lossless CM streams. During the prototyping and demonstration phases of a multimedia testbed [HRKHS96], we noticed that missing a few media units does not result in considerable user discontent, provided that not too many media units are missed consecutively, and such misses occur infrequently. We also noticed that CM streams could drift in and out of synchronization without noticeable user dissatisfaction. Based on these observations, we were inspired to investigate the perceptual tolerance to discontinuity caused by media losses and repetitions, and to that of varying degrees of missynchronization across streams. Following the methodology of [Ste96], we designed a mathematical model and metrics to measure stream continuity and synchronization in the presence of media losses [WS96]. This paper reports the results of a user study to validate those metrics, and consequently, quantify human tolerance of transient continuity and synchronization losses with respect to audio and video.

This study yielded a number of interesting observations concerning the human perception of the quality of CM presentations, of which the main ones are listed below.

- The pattern of user sensitivity varies, depending on the type of defect.
- Viewer discontent for aggregate video losses gradually increases with the amount of loss.
- For other types of losses and missynchronizations, there is initially a sharp rise in user discontent up to a certain value of the defect, and then the discontent plateaus.

* This work is supported by Air Force contract number F30602-96-C-0130 to Honeywell Inc, via subcontract number B09030541/AF to the University of Minnesota, and DOD MURI grant DAAH04-96-10341 to Cornell University
Correspondence to: J. Srivastava

– Rate fluctuations fall somewhere in between, and our experiments indicate that humans are much more sensitive to audio losses than to video losses.
– At a video playout rate of 30 frames per second, average loss below 17/100 is imperceptible, between 17/100 and 23/100 is tolerated, and above 23/100 is unacceptable.
– While video content is always continuous, i.e., there is always some picture on the screen, audio content can be continuous or bursty. Music is continuous, while speech is bursty, i.e., there are talk-spurts interspersed with periods of silence. Any experiment on audio continuity must account for this. We did not consider this *a priori*, and hence ended up mostly eliminating silence from the audio. The only observation we have in this regard is that an average of 21/100 silence elimination does not result in user discontent. However, this issue needs to be studied in much greater detail.
– Consecutive video loss of two video frames in 100 does not cause user dissatisfaction. However, losing two consecutive video frames is noticed by most users, and once this threshold is reached there is not much room for quality degradation due to consecutive losses.
– Consecutive loss of up to three frames was unnoticeable for audio.
– Humans are not very sensitive to video rate variations, in contrast to the high degree of sensitivity to audio. Our results indicate that even a 20% rate variation in a newscast-type video does not result in significant user dissatisfaction. The results with audio rate variations is quite different. Even about 5% rate variation in audio is noticed by most observers.
– Momentary rate variation in the audio stream seemed amusing for a short time, but it soon resulted in being considered an annoyance, and participants concentrated more on the defect than the audio content.
– At aggregate audio-video synchronization loss of about 20/100, human tolerance plateaus. This figure is about three frames for consecutive audio-video synchronization loss. These results are consistent with the findings of [Ste96], where a constant missynchronization of about 120 ms is noticed but accepted by most participants, but about 200 ms constant missynchronization is considered an annoyance.

Our results can be used by algorithm designers in two ways. Firstly, given a level of consumer satisfaction, they can be used to compute the maximum permissible defect of each type. Secondly, in a situation where avoidance of all types of defects is not possible, the tabulated results can be used to choose to sustain one kind of defect over any other, that results in minimal user discontent.

The rest of the paper is organized as follows. Section 2 describes our metrics for continuity and synchronization. Section 3 describes the experimental setup and methodology. Sections 4 through 7 analyze experimental results. Finally, Sect. 8 describes overall conclusions that can be drawn from our experiments, potential use of the results, and our ongoing work in this area. Section 9 contains a concluding summary.

2 Metrics for continuous media

This section summarizes the continuity and synchronization metrics used in our experiments, details of which are provided in [WS96].

2.1 Metrics for continuity

Continuity of a CM stream is metrized by three components; namely *rate*, *drift* and *content*. The ideal rate of flow and the maximum permissible deviation from it constitute our *rate* parameters. Given the ideal rate and the beginning time of a CM stream, there is an ideal time for a given LDU to arrive/be displayed. For the purposes of describing these metrics, envision the evolution of a CM stream as a train of slots with successive slot numbers, where each slot can be filled with some unit of data, such as a video frame (referred to as logical data units – LDUs in the uniform framework of [SB96]). In a *perfect* stream, these LDUs will appear at the beginning of the *slot time*, and, in a lossless stream, there is an ideal sequence of LDUs to appear in a slot: i.e., the i^{th} slot should contain the i^{th} LDU. Given the non-determinism that exists in systems, the i^{th} LDU may not appear in the i^{th} slot. This results in sequencing losses, measured in terms of *aggregate loss factor (ALF)* and *consecutive loss factor (CLF)*. Also, due to timing delays, the LDUs may not appear at the beginning of their slot time. This results in timing deviations measured in terms of *aggregate drift factor (ADF)* and *consecutive drift factor (CDF)*.

In order to define losses, we define *unit sequencing loss* (USL). To define unit sequencing loss, envision a CM stream as a train of slots with successive slot numbers, as given in Fig. 1. Some slots may be filled with LDUs. We define a USL only for slots that are non-empty, i.e., they are filled with some LDU. Suppose $s(k)$ is the LDU at slot $s((i))$ of stream $s(\cdot)$. Suppose the immediately previous non-empty slot to slot $s((i))$ is slot $s((i-l))$, where $l > 0$, and it is occupied by LDU $s(j)$. In case there are no skips, repeats or misses, if slot $s((i))$ is occupied by LDU $s(k)$, then slot $s((i-l))$ should be occupied by LDU $s(k-l)$. Hence, the USL incurred at slot $s((i))$ due to skips and repeats is $\|k-l-j\|$. The USL due to missing LDU at slot $s((i))$ is $(l-1)$, precisely because there are $(l-1)$ empty slots in between slots $s((i))$ and $s((i-l))$. Hence the maximum of sequencing loss due to skips, repeats and misses at slot $s((i))$, say $USL(i)$, is $\max\{\|k-l-j\|, l-1\}$. Consequently, we define $\max\{\|k-l-j\|, l-1\}$ to be the *USL* at slot $s((i))$. In order to measure the sequencing loss at the beginning of a stream, we assume that every stream has a hypothetical slot $s((-1))$ with number -1, containing a hypothetical media granule $s(-1)$.

Now, we use USLs to specify sequencing profiles. Our sequencing profile specifies allowable average and bursty USLs, which are specified by the ALF and the CLF.

An ALF of n/m for a stream means that n is the sum of USLs allowed within any window of m successive slots for LDUs, i.e., $\max\{\sum_{k=i}^{i+m}\{USL(k) : USL(k) \neq \perp\} \leq n$ for any $i \geq 1$. The CLF is the maximum sum of non-zero CLFs, i.e., $\max\ \{\sum_{k=i}^{i+l}\{USL(k) : USL(k) \neq \perp,\ \forall k\ (i \leq k \leq i+l)\} : i, l \geq 1\} \leq CLF$.

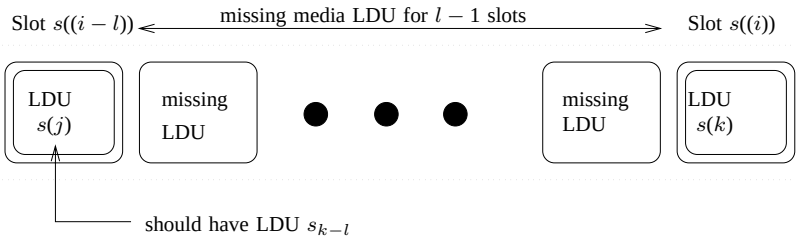


Fig. 1. Unit sequencing loss

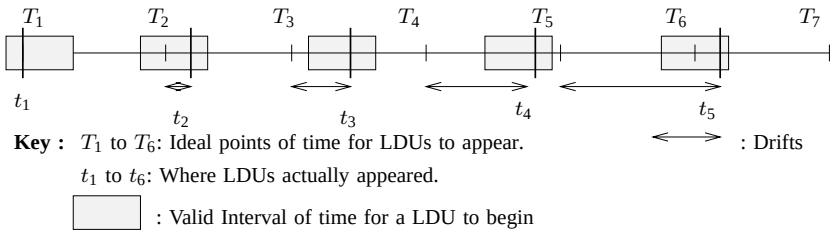


Fig. 2. Drifts in a stream

Definition (rate profile): Let $s(i)$, $s(i+1)$, ..., etc. be successive slots for a stream $s(\cdot)$, where the starting time for slot $s(i)$ is t_i. Stream $s(\cdot)$ is defined to have a rate profile (ρ, σ) iff $\forall i\ t_{i+1} \in [t_i + \frac{1}{\rho+\sigma}, t_i + \frac{1}{\rho-\sigma}]$.

Definition (unit granule drift): Unit granule drift at slot $s(i)$, $UGD(i)$, is defined as the time difference between the nominal start time, T_i, of slot $s(i)$ and its actual starting time, t_i, i.e., $UGD(i) = \|t_i - T_i\|$.

Figure 2 shows examples of these concepts. If the LDU $s(j)$ is omitted, then t_j is undefined, i.e., $\perp$, and hence $\|t_j - T_j\|$ and $UGD(j)$ are undefined. Using the sequence of UGD's $\{UGD(i) : i \geq 1\}$, we can define the drift profile (ADF, CDF). An ADF of $d//m$ means that no consecutive m granules can have a sum of more than d time units of granule drift, i.e., $\sum_{k=i}^{i+m}\{UGD(k) : UGD(k) \neq \perp\} \leq d$ for any $i \geq 1$. A CDF of d' means that the sum of consecutive non zero delays can be at most d' time units, i.e., $\max\{\sum_{k=i}^{i+l}\{UGD(k) : UGD(k) > 0\ \forall k\ (i \leq k \leq i+l)\} : i, l \geq 1\} \leq d'$.

For example, the first four LDUs of two example streams with their expected and actual times of appearance, are shown in Fig. 3. In the first stream, the LDU drifts are 0.0, 0.2, 0.2 and 0.2 s. Accordingly, the stream has an aggregate drift of 1.2 s per 4 time slots, and a non-zero consecutive drift of 1.2 s. In the second stream, the largest consecutive non-zero drift is 0.2 s and the aggregate drift is 0.3 s per four time slots. The reason for a lower consecutive drift in stream 2 is that the unit drifts in it are more spread out than those in the first stream.

2.2 Metrics for synchronization

For a group of synchronized streams, there is a natural collection of LDUs that must be played out simultaneously. The largest difference in the LDU numbers between any two pairs in such a group is the unit synchronization loss. The aggregate and largest non-zero consecutive unit synchronization loss is referred to as *aggregate synchronization content loss (ASL)* and *consecutive synchronization content loss (CSL)*, respectively. In the example of Fig. 3, due to losses of LDUs, there are unit synchronization content losses at the first and the last pairs of LDUs, resulting in an ASL of 2/4 and a CSL of 1.

In a perfectly synchronized collection of streams, the i^{th} LDU of each stream should start playing out at the same instant of time. Failure to accomplish this ideal is measured by the maximum difference between the display start time of the LDUs in the group, and is referred to as the *unit synchronization drift (USD)*. The aggregate of USD's over a given number of LDU slots is the aggregate synchronization drift, and the maximum of such non-zero consecutive USD's is the consecutive synchronization drift. They measure the average and bursty time drifts in synchronization. In Fig. 3, the two streams have USDs of 0.2, 0.2, 0.0, and 0.4 s, respectively, resulting in an aggregate synchronization drift of 0.7/4 s, and a consecutive synchronization drift of 0.4 s.

Playout rate of a collection of synchronized streams is determined by the rates of component streams. The rate variation of a collection of synchronized streams is the maximum difference between the fastest and slowest rates.

2.3 Relationship between metrics

Two types of specifications must be satisfied in a synchronized rendition of a collection of CM streams. They are, synchronization parameters of a collection of streams and continuity parameters of their components. In the current section, we state the definability of some of these parameters with respect to others. We show the following facts, and their stated consequences follow. Proofs of these facts appear in [WS96].

1. Mixing profiles of a collection of synchronized streams cannot be defined in terms of stream parameters of their components.
   Consequence. It is not possible to control the mixture of samples displayed together only by exercising control over individual streams, without having a mechanism to handle cross-stream effects.
2. Rate profiles of a collection of synchronized streams can be defined in terms of rate profiles of their components.
   Consequence. The rate of a synchronized rendition can

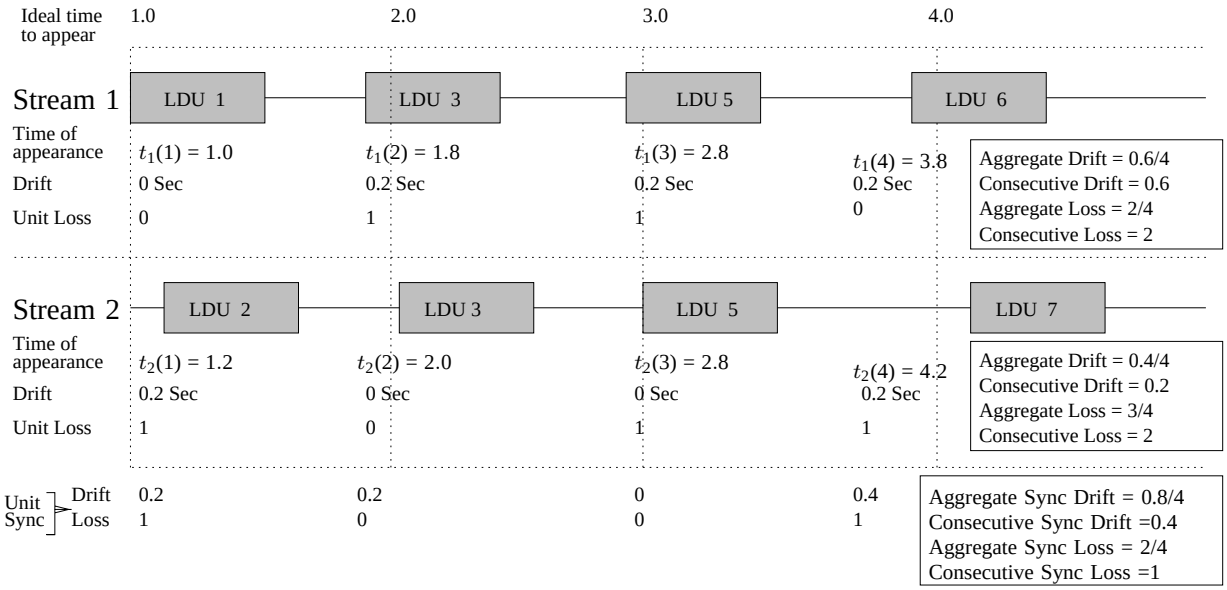


Fig. 3. Two example streams used to explain metrics

be controlled by controlling rendition rates of its component streams.

3. Except for the perfect case, the synchronization drift profile of a collection of streams is not definable in terms of the drift profiles of its components, although the aggregate synchronization drifts can be bounded by drift profiles of component streams.
   Consequence. It is possible to control average timing drifts in a synchronized rendition by controlling timing drifts of its component streams.
4. Consecutive synchronization drift of a collection of synchronized streams is not definable in terms of the drift profiles of its component streams.
   Consequence. It is not possible to control bursty timing drifts between a collection of synchronized streams by controlling the individual timing drifts of its component streams.

3 Experimental design

Originally, we planned to measure and validate the tolerable ranges of all our metrics. Due to the inability to control timing precisely on computers, we decided to use professionally edited pre-recorded segments of audio and video. Even the professional editing equipment was unable to control the appearance of video and corresponding audio to millisecond time granularity. Hence, we focused only on testing for content and rate parameters. Recent work by [SNL95] has shown that this can be done using a specialized hardware board attached to a Sun workstation. We have received this board from Schmidt et al., and plan to use it in our future work.

Our experiments consisted of eight sets; aggregate and consecutive content losses of audio, video and synchronization were six of them, and rate variations in the audio and video streams were the remaining two. Of the eight sets, three consisted of audio-only segments, another three consisted of video-only segments, and the remaining two consisted of audio and video segments.

3.1 Design concerns and the pilot study

Several issues in survey design and psychological evaluations arise in carrying out user evaluation of human perception. A key issue is the proper design of the evaluation instrument, i.e., the survey form, so that the data collected can be used to test the hypotheses; minimizing the effects of extraneous variables and participant bias, and avoiding conveying designer bias to the participants. In our design, we have strived to achieve these goals. In designing our experiment, the experimental methodology and techniques of analysis used in [Ste96, AFKN94] have been useful to us.

In order to evaluate potential suitability of our experimental methodology and design to the intended task, we conducted a pilot study with about ten participants. The results of this study and professional help [Molly Fiedler, personal communication] made us change many things in the questionnaire, video tapes, and the environment in which the experiment was carried out. In the tape, we decided to first show a clip in its perfect form. This helps each participant establish a baseline against which to evaluate the quality of the other tapes. This was essential due to the fact that TV and broadcast media that our participants are most familiar with do not usually have the kind of defects that we wanted observed. We provided a potential list of defects, some of which were not in our clips. This was done because many participants do not use the same words to describe a defect, and an *ipso facto* defect categorization leads to too many categories. Clips with varying amounts of defects of the same type were grouped together, with a clip having no error included in the group. Each experiment containing audio, video or both was identified as such, to ensure that the absence of either media type not be considered a defect.

In the design of the survey, we had to make substantial changes after the pilot study. It was determined that the sur-

Fig. 4. Shots of audio-visual clips used in the experiment

vey should have a tabular format, as opposed to having a page per clip. The sheer size of survey forms seems to discourage some potential participants. The order and wording of questions must be changed to suit an average American college student audience. We also decided not to allow individuals to take the survey on their own, so that the environment of the presentation, and answers to participant doubts and questions during the experimental runs can be controlled. The Likert scale was changed from [1, 8] to [1, 10], where 1 was poor and 10 was excellent. We also asked the participants to categorize each clip as *Do not mind the defect if there is one*, *I dislike it and it's annoying*, and *I am not sure*, similar to the survey in [Ste96].

3.2 Design decisions

Audio-video segments of 30 s duration were taken from a bust view of two articulate speakers (Fig. 4), with no particular accents, describing neutral subjects. The chosen speakers were unknown to participants in the study. This was done to avoid any biases that may carry over about the speakers into our study. Neutral accents were chosen to avoid any misinterpretation of words in the face of introduced defects, and also to give our participants the benefit of listening to a voice that comes with the most familiar pronunciation. The contents used by the two speakers were (a) the care they take in organizing their lectures, and (b) the concentration spans of junior high school students. None of our participants were teachers, nor junior high school students. The length of test segments were chosen to be 20–30 s, since, according to [Ste96], about 20 s suffices for participants in an MM user study to form their opinions about a clip. Although the head view results in the most number of defects being perceived [Ste96], we chose the bust view, because it represents the news media type of a situation better than a talking head occupying an entire screen.

3.3 Parameters used in the experiments

The tapes were made with the following characteristics. In the aggregate media loss experiments, the consecutive losses were kept constant at three video frames, under the normal speed of 30 frames per second. The media losses were created by introducing *jump cuts* in the NTSC time code. For the rate variation experiment, a nominal rate of 30 frames per second rate was maintained, but a square sinusoidal wave, with each quarter wave lasting 5–6 s was produced. For the ASL experiment the CSL was fixed at four video frames at 30 frames/second. For the CSL experiment the aggregate synchronization loss was fixed at 40/100. The master tape consisted of an introductory part lasting about 3 min, after which the two perfect clips were shown, followed by three groups of experiments: video, audio and synchronization. Within each group, the sub-group order was aggregate loss, consecutive loss and rate variation experiments. Within each experiment, defective clips were arranged in the random order given in Table 1. For each experiment there were about five to six clips, with varying degrees of controlled defects, that were shown in random order.

3.4 Administering the experiment

Experiments were conducted in small groups of 3–6 participants, for a total of 70 participants, chosen mostly from students at the University of Minnesota, who participated in our study voluntarily. In order to draw participant attention to potential defects, the background noise was kept to a minimum and the contents of clips were deliberately selected to be not too engrossing. We also told the participants that the objective of our study was to look for defects, and provided a sample list of them. At the beginning of the survey, we showed the two clips in their perfect form. As expected, most participants found the endeavor boring and very repetitive, although a fair number found some clips to be rather amusing. For every group of participants, all eight experiments were conducted in one sitting that lasted about 45 min. After each clip was shown, the participants were asked to fill out the corresponding row of scores in a survey form. The

Table 1. Order of defects in test clips

Experiment	Media	Defect in test clips					
Aggregate loss	Video	6/100	21/100	12/100	3/100	0/100	
Consecutive loss	Video	0	1	5	4	3	2
Rate variation	Video	10%	0%	02%	20%	15%	6%
Aggregate loss	Audio	6/100	21/100	12/100	3/100	0/100	
Consecutive loss	Audio	0	1	5	4	3	2
Rate variation	Audio	10%	0%	02%	20%	15%	6%
Aggregate synchronization loss	A/V	40/100	4/100	16/100	24/100	0/100	
Consecutive synchronization loss	A/V	15	3	10	0	5	20

sample survey used for the first clip is given in Fig. 5. The survey consists of an introductory description, six tables (one per experiment) and a questionnaire about the participant's experience with TV production. As seen from the sample table given in Fig. 5, each participant had to grade each clip on a Likert scale [Opp83] from 1 to 10, identify defects perceived, and state if the defect was annoying, not so, or could not decide, which we call the *acceptability score*.

3.5 Processing the surveys

The results of the surveys were entered into a database, and visualized using Matlab [PESMI96]. As expected, increase in defects resulted in a decrease of user satisfaction, except for the experiment on aggregate losses of audio. The data as taken from the surveys, the average and standard deviations of Likert values, and the ratio of participants who considered the clip to be perfect, acceptable and unacceptable, were plotted for each experiment. These graphs were smoothed by using a cubic spline interpolation provided by Matlab. The analysis of the data and conclusions drawn from them follow in Sects. 4 through 7.

Two remarkable trends emerge from our results. First is that, for some kinds of defects, there is a gradual increase in user discontent with increasing defects. Aggregate video loss is a clear example of this kind. Second is that, for some other kinds of defects, there is a sharp increase in user discontent that plateaus after a specific threshold. Synchronization and consecutive loss are clear examples of this kind. Rate fluctuations are somewhere in between, and humans seemed to be far less tolerant of audio rate fluctuations than of video.

4 Aggregate loss experiment for media streams

There were five clips with aggregate media losses ranging from 3/100 to 21/100, with a consecutive loss factor of 3 LDUs. The presentation order of these clips is given in Table 1. For the aggregate loss experiment of video streams, as evident from data tabulated in Fig. 6b and visualized in Fig. 6a, as the aggregate media loss increases the distribution of Likert values shifts from the higher end towards the lower end of the spectrum. The values on the vertical axis are the acceptability scores for the experiments. This trend indicates that increased aggregate video loss leads to increased viewer discontent.

We were expecting the same trend in the corresponding experiment on audio, but as observed from data tabulated in Fig. 6d and visualized in Fig. 6c, our expectations were not fulfilled to the same extent as for video. A closer examination of our tapes revealed that most eliminated LDUs in the audio stream correspond to silence. Consequently, although it requires further experiments to justify our speculation about aggregate audio drops, current results indicate that aggregate silence elimination in the audio stream does not result in considerable user discontent in the range 0/100–21/100. We speculate that further silence elimination would reach a point of considerable listener discontent, as the speech will appear unnaturally hurried. Notice that the higher end Likert scales of Fig. 6D provide evidence in support of this trend. Our ongoing work includes further experimentation to test this speculation. Silence elimination can be used very profitably by computer system designers to reduce resource requirements, since it requires no processing, transmission, storage, etc.

To further our understanding of the pattern of user discontent, we plotted the average and standard deviations of Likert values against the losses for video and audio, given in Fig. 7a and c, respectively, which clearly brings out the trend. The lower standard deviation at the higher values of the average Likert scale indicates that there is higher consensus in the judgment expressed by its mean. Also notice that the maximum standard deviation is about 2, a reasonable 1/5 of the total score.

The acceptability scale, plotted in Fig. 7b and d, respectively, shows the regions in which users expressed clear unacceptance, willingness to tolerate, and perfect acceptance. In all our graphs, we notice a correlation between the average Likert value in the Likert scale and the curve that separates the *unacceptable* region from the rest. This seems to indicate that the two metrics that were used in the two other reported user studies in multimedia [Ste96, AFKN94], namely the Likert and the acceptability scales, have a strong relationship to each other, and consequently can be used in our type of study interchangeably.

If the Likert and acceptability scores are graphed together, the former intersects the latter at about 17/100 aggregate media loss, and the unacceptability curve at about 23/100 media loss. Modulo our experimental results, these observations imply that 17/100–23/100 is the noticeable but tolerable region for aggregate video losses. Similar analysis applied to the results of the audio experiment yields that, within our operational range, i.e., 0/100–21/100, aggregate audio losses went unnoticed.

Experiments with Video Only Clips

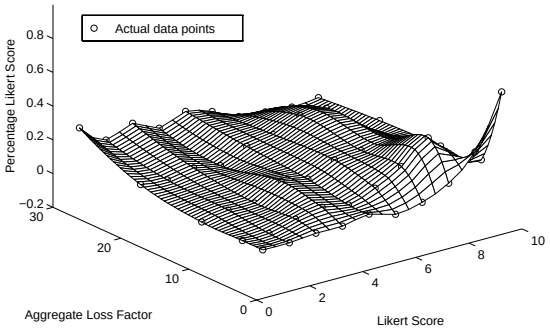
These experiments have NO SOUND. Please watch the silent video and fill out the following tables.

Clip Number	Grade the quality of the clip	Did you notice a defect ? If so, please describe it	If your TV programs had this error how would you categorize it?		
	1 (poor) to 10 (excellent)	i.e., skip, stutter breaks, missynchronization, gaps distortions etc.	I don't mind the defect	I dislike it. it's annoying	I am not sure It depends
Group 1 Clip 1	1 2 3 4 5 6 7 8 9 10				
Clip 2	1 2 3 4 5 6 7 8 9 10				
Clip 3	1 2 3 4 5 6 7 8 9 10				
Clip 4	1 2 3 4 5 6 7 8 9 10				
Clip 5	1 2 3 4 5 6 7 8 9 10				

Fig. 5. A sample table from a blank survey form

Video Experiment

Likert Value Distribution in Aggeregate Video Loss Experiment



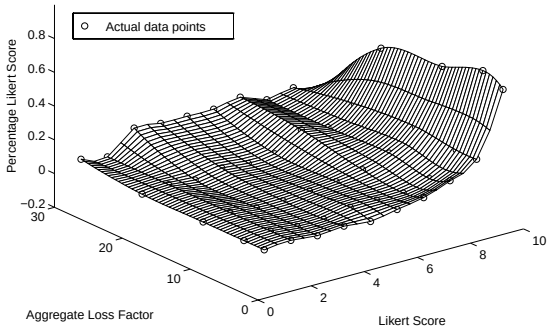
a

Agg Loss	Number of Likert Scores out of a total of 70									
	1	2	3	4	5	6	7	8	9	10
3/100	0	0	1	1	3	0	2	7	17	39
6/100	0	0	3	5	3	7	13	22	10	7
12/100	0	2	2	8	6	11	12	14	10	5
21/100	4	3	5	3	8	10	15	15	3	4
30/100	16	8	12	7	11	8	3	2	1	2

b

Audio Experiment

Likert Value Distribution in Aggeregate Audio Loss Experiment



c

Agg Loss	Number of Likert Scores out of a total of 70									
	1	2	3	4	5	6	7	8	9	10
3	0	1	0	1	0	2	4	8	14	40
6	0	1	0	0	2	2	2	5	14	44
12	0	1	0	0	1	2	4	8	16	38
21	0	1	0	2	1	2	9	7	14	34
30	3	1	10	9	9	9	11	7	9	2

d

Fig. 6a–d. Data from the aggregate loss factor experiment

5 Consecutive loss experiment for media streams

There are six clips with aggregate media losses ranging from 0 to 10 consecutive LDUs, with the presentation order of clips as shown in Table 1. As seen from results tabulated in Fig. 8b and d, and visualized in Fig. 8a and c, increasing consecutive loss results in a sharp rise in viewer discontent. This is evidenced by the concentration of lower Likert values around 3–5 consecutive media losses in data from both video and audio streams, as given in Fig. 8b and d, respectively.

This trend is further illustrated by the average Likert and acceptability graphs shown in Fig. 9a,c and Fig. 9b,d, respectively. As seen in Fig. 9d, for audio streams three to four consecutive frame losses receive a Likert score of 9. For video, as seen from Fig. 9b this limit is two frames. Compared with the video aggregate loss experiments shown in Fig. 7, acceptability scores have a thin margin for noticeable-but-tolerable consecutive losses, although the margin for video losses is slightly higher than those for audio. In contrast to average video losses, graphed in Fig. 7b, user discontent with consecutive losses sharply rises and then plateaus at two and three frames for video and audio, respectively. Standard deviation for acceptability values for both media, as shown in Fig. 9a and c is approximately 2 units. At the high end of the scale, the standard deviation for the video stream is lower, indicating more consensus in the rating. Because of the thin margin for the acceptable region, the intersection of Likert graphs and acceptability graphs remain single values, i.e., 1 and 2 for video and audio, respectively.

6 Rate variation experiment

There were six clips with 0–20% rate variation from an average rate of 30 frames/second, following a pattern of a square sine curve of five quarter-frame lengths. The presentation order of these clips is as shown in Table 1. As evident from data tabulated in Fig. 10b,d and visualized in Fig. 10a,c, user

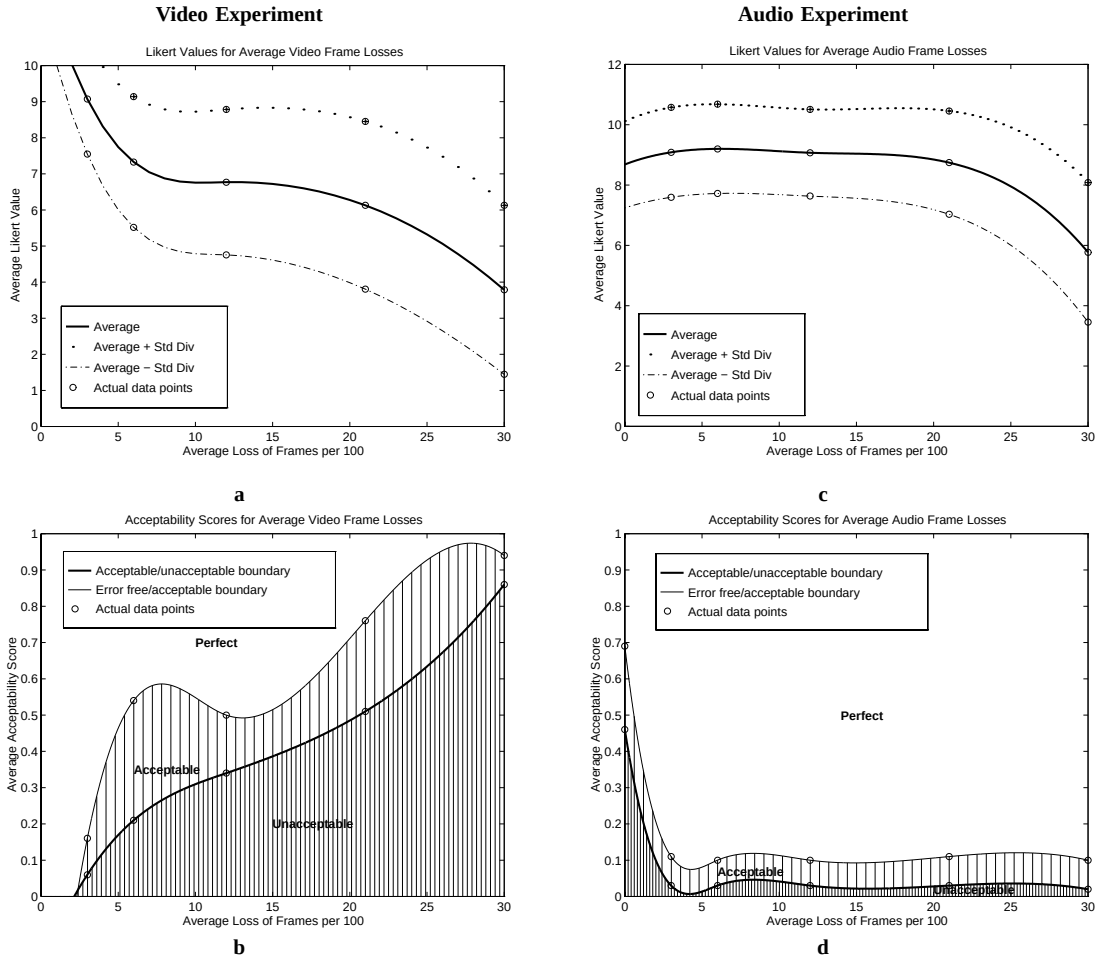
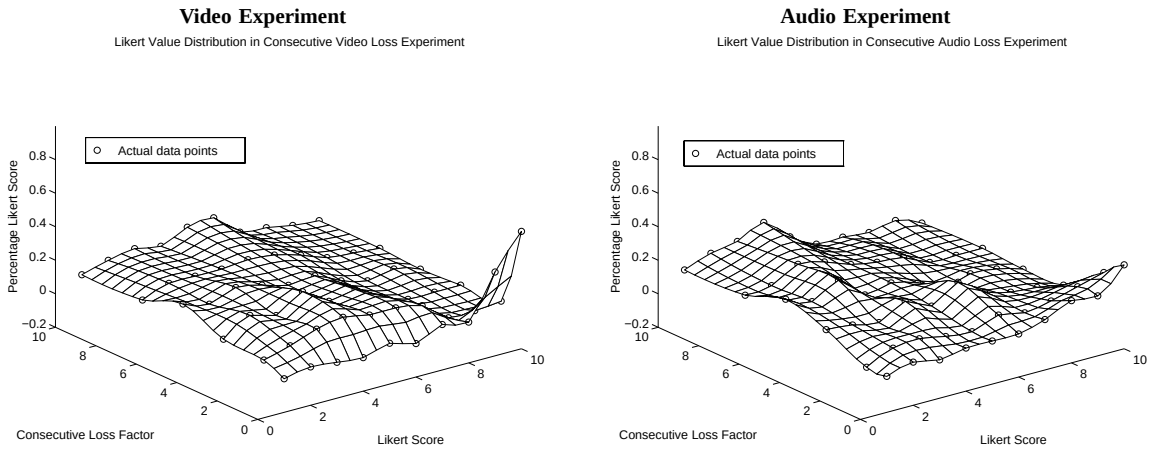


Fig. 7a–d. Summarized results of the aggregate loss factor experiment



Consec. Loss	Number of Likert Scores out of a total of 70									
	1	2	3	4	5	6	7	8	9	10
0	0	2	1	0	3	0	5	3	21	35
1	4	7	11	13	11	11	9	1	1	2
3	5	12	9	16	9	11	5	1	1	1
5	12	9	13	11	8	9	4	1	3	0
7	6	10	10	10	11	10	7	3	2	1
10	5	8	10	8	13	14	5	4	2	1

b

Consec. Loss	Number of Likert Scores out of a total of 70									
	1	2	3	4	5	6	7	8	9	10
0	1	4	2	4	4	4	7	12	11	21
1	1	6	7	5	4	8	6	8	11	14
3	9	8	6	12	7	14	7	4	3	1
5	14	10	16	7	9	6	1	4	3	0
7	8	9	15	13	9	7	6	1	1	1
10	7	11	13	18	9	3	3	2	4	0

d

Fig. 8a–d. Data from the consecutive loss factor experiment

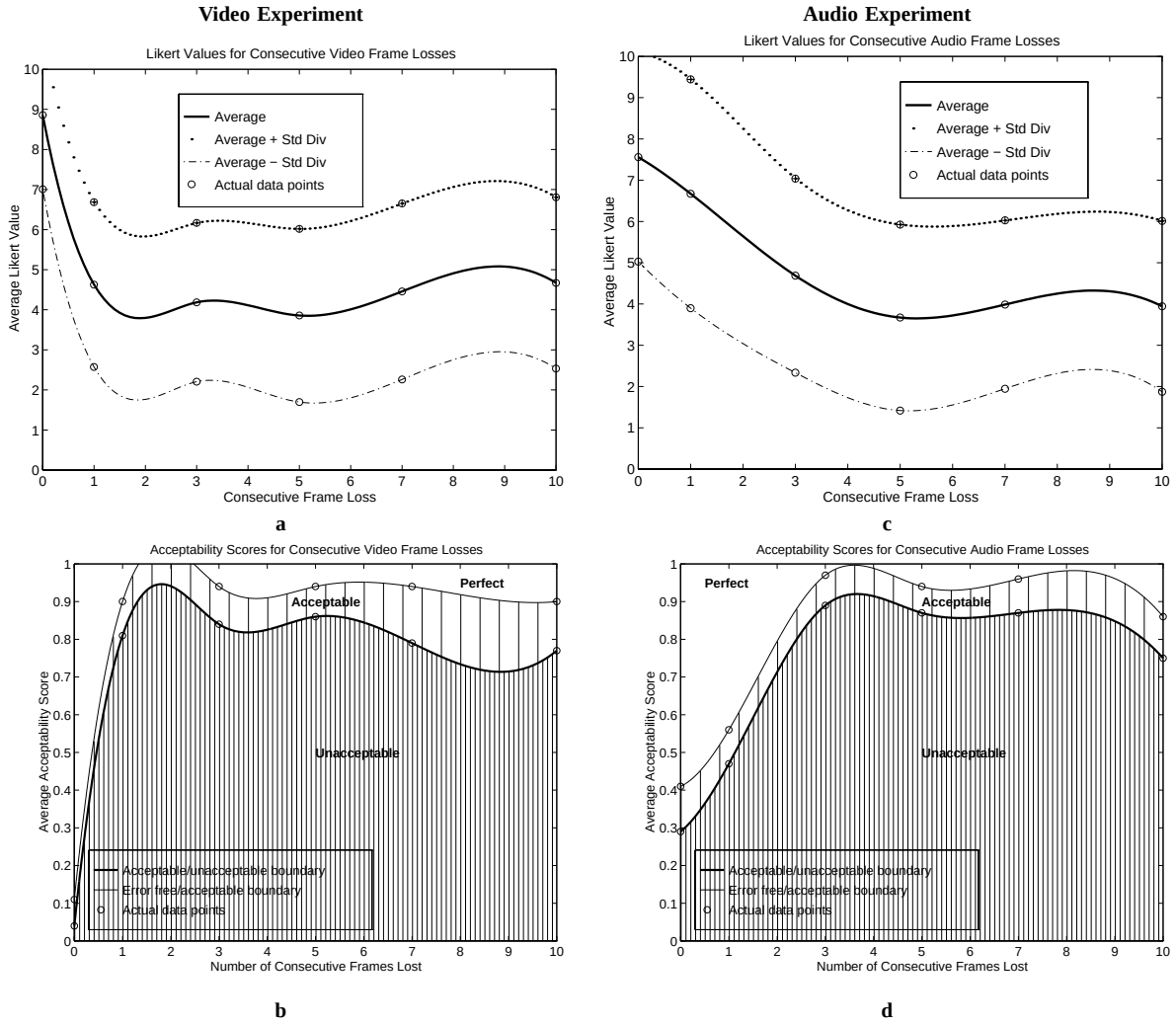


Fig. 9a–d. Summarized results of the consecutive loss factor experiment

discontent shifts from the higher end to the lower end of the Likert scale with an increase in the amplitude of the sine wave rate, indicating that increasing rate fluctuations lead to increased viewer discontent.

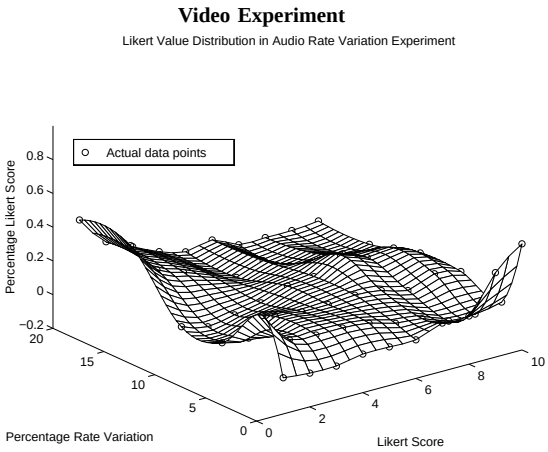
To further our understanding of the pattern of user discontent, we tabulated the average and standard deviations of Likert values against the losses, as given in Fig. 11a,c, which clearly brings out the trend. The lower standard deviation at the higher values of the average Likert scale indicates that there is higher consensus in the judgment expressed by its mean. Also, the maximum standard deviation in Fig. 11a,c is about 2. Notice that the average Likert value in the audio case decreases more uniformly, compared to video. This trend implies that we are not very sensitive to the rate fluctuations in video, as compared to those in audio. Further, audio has a uniformly lower score on the Likert scale than video, further substantiating this claim. Data on acceptability scores has been plotted in Fig. 11b and d, and shows the corresponding plateaus and trends similar to those in average Likert scales.

If the Likert and acceptability scores are graphed together, the former intersects the latter for audio at about 7–8%. These results imply that up to about 20% of video and 7% of audio rate variations are tolerated and, after about 8%, audio rate variations become intolerable. In this experiment, two metrics, namely average Likert values and average acceptability scores, show a strong positive correlation.

7 Transient synchronization loss experiments

As stated, there are six clips each for aggregate and synchronization loss experiments. In the aggregate loss experiment they range from 0/100 to 40/100 with a constant consecutive loss of 4, and in the consecutive losses experiment they range from 0 to 20 with an aggregate synchronization loss of 40/100. The presentation order of these clips was as given in Table 1. For synchronization loss experiments, as evident from tabulated data in Fig. 12b and d, and visualized in Fig. 12a and c, as the losses increase, the distribution of Likert values shifts from the higher end to the lower end of the scale, indicating that increased transient synchronization losses lead to increased viewer discontent.

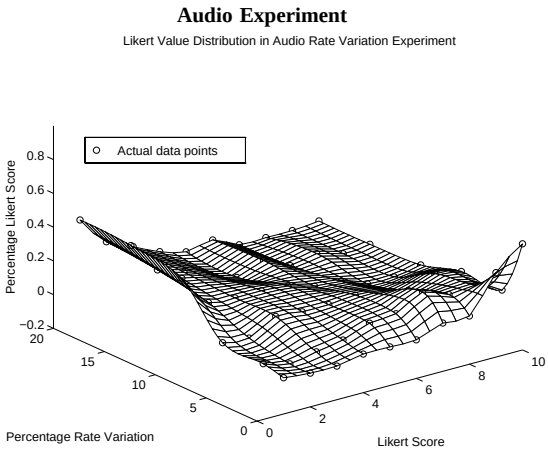
To further our understanding of the pattern of user discontent, we tabulated the average and standard deviations of Likert values against the losses in Fig. 13a and c, which clearly illustrate the trend of average Likert score decreasing



a

% Rate Variation	Number of Likert Scores out of a total of 70									
	1	2	3	4	5	6	7	8	9	10
0	2	0	1	1	2	1	5	10	18	30
2	0	1	4	4	7	11	11	10	10	12
6	0	1	1	1	8	6	6	13	20	14
10	0	1	2	2	6	5	10	11	17	16
15	1	1	4	2	7	8	8	12	13	14
20	0	2	5	5	6	12	9	13	9	9

b

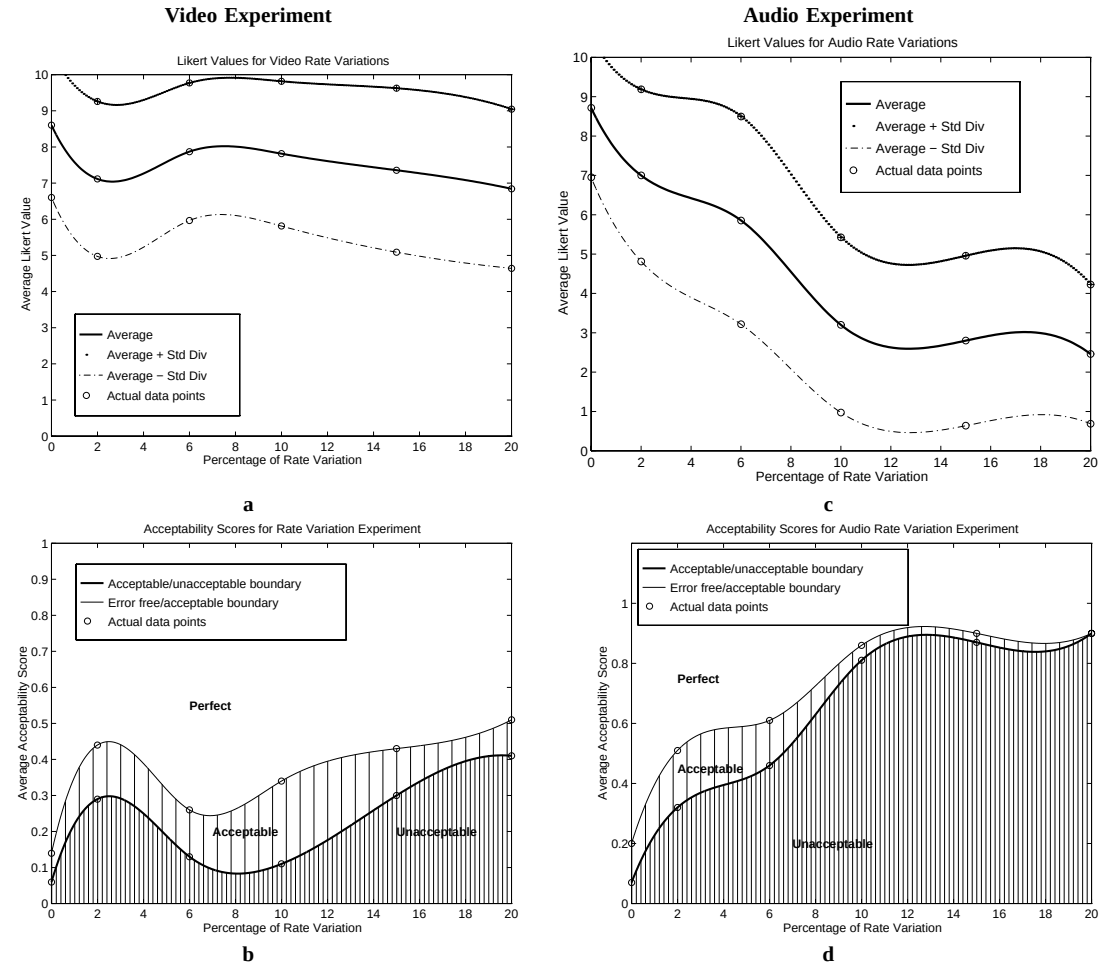


c

% Rate Variation	Number of Likert Scores out of a total of 70									
	1	2	3	4	5	6	7	8	9	10
0	1	0	0	2	2	2	6	6	21	30
2	3	0	2	4	8	9	15	11	12	7
6	4	3	9	8	8	9	8	5	9	7
10	23	10	9	9	8	5	4	0	0	2
15	27	14	10	6	2	6	1	3	0	1
20	28	16	11	6	3	5	0	0	0	1

d

Fig. 10a–d. Data from rate change experiment



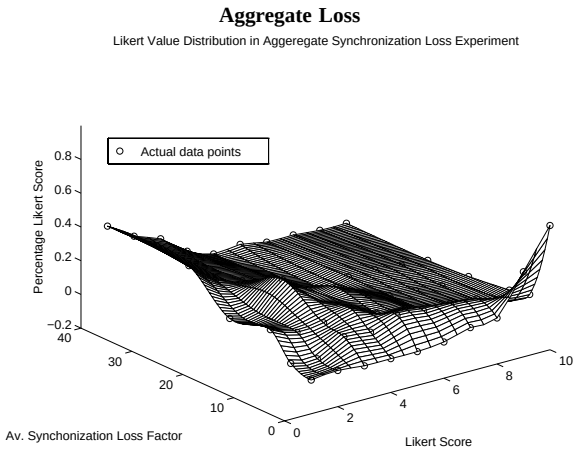
a

c

b

d

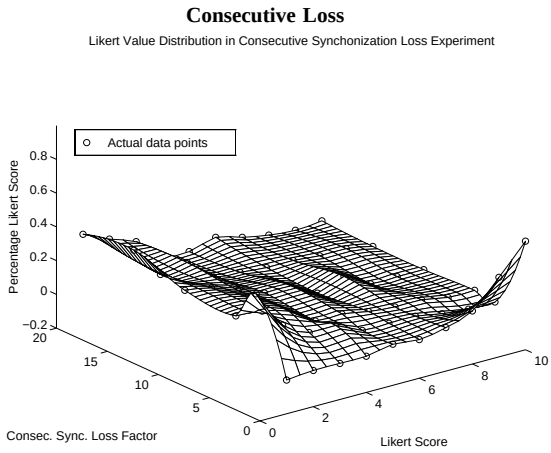
Fig. 11a–d. Summarized results of the fluctuating rates experiment



a

Agg Loss	Number of Likert Scores out of a total of 70									
	1	2	3	4	5	6	7	8	9	10
0	0	1	0	0	0	1	4	5	21	37
4	3	2	8	7	11	10	10	9	6	5
8	13	9	10	13	8	8	3	2	1	3
16	10	10	20	9	11	6	2	1	0	1
24	24	15	12	10	4	0	2	2	1	0
40	25	18	14	6	2	3	1	1	0	0

b

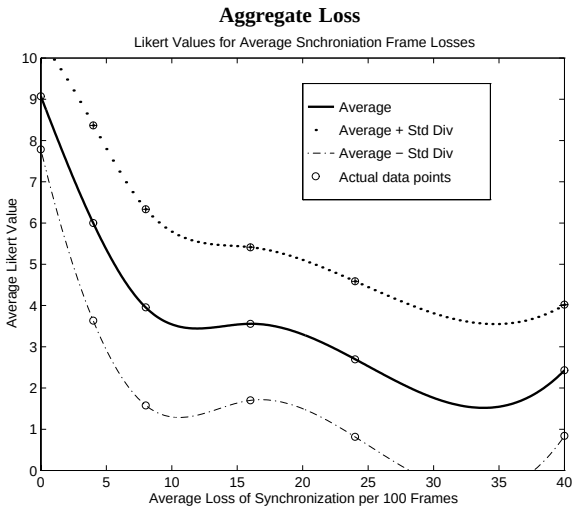


c

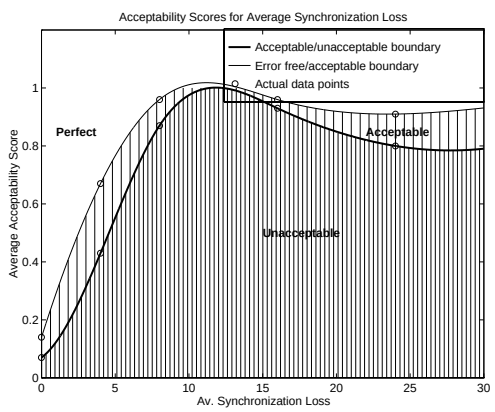
Consec. Loss	Number of Likert Scores out of a total of 70									
	1	2	3	4	5	6	7	8	9	10
0	0	1	1	1	3	2	4	8	19	31
3	30	14	8	8	3	3	3	1	0	0
5	17	16	13	6	6	4	3	3	1	1
10	18	16	12	8	8	2	6	0	0	0
15	25	12	14	5	4	3	3	2	2	0
20	22	17	13	4	3	6	3	1	0	1

d

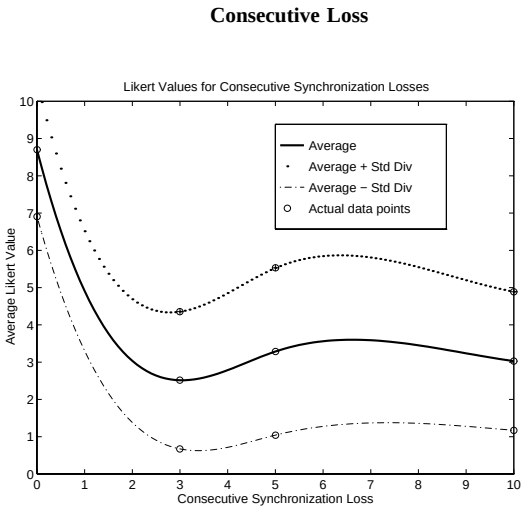
Fig. 12a–d. Data from synchronization loss experiments



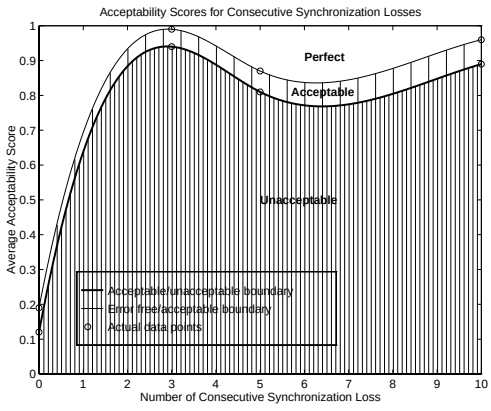
a



b



c



d

Fig. 13a–d. Summarized results of synchronization loss experiments

with increased synchronization losses. As in the case of consecutive media loss experiments, there is a sharp increase in the acceptability values which plateaus around 12/100 and 3 for average and consecutive losses respectively.

The acceptability scale, visualized in Fig. 13b and d, shows the regions in which users expressed clear intolerance, willingness to tolerate, and perfect acceptance. This scale also sharply decreases and plateaus at 12/100 and 3 for average and consecutive losses. These correspond to the peaks in the two figures.

The intersections for average Likert and acceptability curves indicate that 6/100-7/100 is the range for tolerable average synchronization losses, and a single frame is the tolerance limit for consecutive synchronization losses.

As in all other graphs, we notice a clear correlation between the average Likert value and the curve that separates the *unacceptable* region from the rest on the acceptability scale, indicating a strong correlation between them in synchronization experiments.

8 Further inferences and usage of experimental results

This section provides some further inferences from our experimental data, their projected usefulness, and our ongoing work in this area.

8.1 Further inference from experimental results

As stated, two remarkable trends emerge from our results. First is that, for some defects, there is a gradual increase in user discontent with increasing defects. Aggregate video loss is a clear example of this kind. Second is that, for some defects, there is a sharp increase of user discontent that plateaus after a specific value. Synchronization and consecutive losses are clear examples of this kind. Rate fluctuations are somewhere in between, and humans seem to be far less tolerant to audio rate fluctuations than to video. Although we generally concur with the synchronization experimental results obtained in [Ste96], based on our observations, we believe that not all QoS experiments are going to result in such clear-cut boundaries for distinguishability, tolerance and unacceptability for QoS metrics, but they gradually decrease throughout a continuous spectrum of values. This trend is clearly evidenced in our aggregate loss experiment for video, and also in the rate experiments of [AFKN94].

In addition to determining the acceptable ranges for some of our QoS parameters, we can also determine their relative importance. For example, we can directly compare the Likert values of aggregate video losses and aggregate synchronization losses to determine the loss ranges where one of them is more crucial than the other. Some of the potential benefits of these comparisons are discussed in Sect. 8.2.

8.2 Use of experimental results

Our findings can be used in multimedia testbed designs in two different ways. First, given a Likert value, or an acceptability score that would characterize the required *degree of user satisfaction*, it is possible to determine the tolerance to a given defect. For example, with a 0.8 Likert value, a video stream can sustain a 20/100 average loss, 1 consecutive loss, and up to 20% rate variation. For the audio stream, these parameters are 30/100 aggregate silence elimination, 0.7 s worth of consecutive sample losses, and about 10% rate variation. For audio-video synchronization, they are about 7/100 aggregate losses and one consecutive loss. For a given level of user satisfaction, the tolerances of a set of defects, such as the media and synchronization losses investigated in the present paper, can be used directly as limiting values for the corresponding defects. For example, for 80% user satisfaction, we may have 20/100 as the maximum permissible aggregate video loss.

Second, in designing algorithms we can assign relative weights to these losses. For example, comparing the average Likert values of video loss with consecutive synchronization loss, it is clear that the unacceptability region for the former is below that of the latter, and therefore dropping video frames on the average is preferable to losing synchronization consecutively. To compute relative weights for different parameters, we may assign them weights proportional to the average of some user preference parameter such as the average of all Likert values assigned for that parameter, which can be achieved for the given testbed. For example, if a designed testbed can only deliver with an aggregate video loss of 10/100, and a consecutive synchronization loss of 5, compute the average of the Likert values over [0, 10/100] for the aggregate video loss and over [0, 6] for the CSL. Suppose that the former is 7 and the latter is 5.5, then assign these as weights of importance during dynamic playout management. A potential usage of such weights is that the parameter that carries the smallest weight in the range of operation can be ignored in order to avoiding defaulting on ones with higher weights.

8.3 Comparison with existing work

Parameters of human tolerance to audio-video and audio-pointer synchronization were obtained in [Ste96]. They were categorized as undetectable, detected-but-tolerable, and intolerable errors. These parameters are for lossless streams. In a CM client-server paradigm, streams may be delivered through a network. At the lower levels of the protocol stack, the network can drop packets, and, in order to recover from loss, some kind of retransmission is necessary. This may induce intolerable delays and jitters in the CM stream. Suppose instead that the application itself allows for a lossy media stream, through some QoS-based loss characteristics of CM streams, then the retransmission may be unnecessary, and, consequently the delay and jitter at the application level, and the bandwidth at the network level can be saved. Our parameters can be used to compute such QoS-based LDU drops at the application level.

Another observation we have is that, in our testbed, audio and video drift in and out of synchronization, as opposed to behaving statically. Granted that, if maximum drifts were within the limits reported in [Ste96], then the static limits stated therein would apply. However, we postulated that, for transient missynchronizations, the participants would be

more forgiving. As the reported data indicates, this is not the case.

[AFKN94] categorizes audio-visual clips as *high* and *low* in audio, video and temporal dimensions, referred to therein as *video classification schemas (VCS)*. They measure the perceptual importance of each dimension in conveying the total message contained in clips across to the intended viewer. For example, sports footage and talk shows are considered high and low in the temporal dimension, respectively. Such a classification, while rich in semantics and its relevance to human perception, requires some extra effort, and the servers need to be enriched to understand their significance. This may mean extra effort by the producers or some other intermediate personnel. In this respect, our test clips should be considered low in the temporal dimension and (perhaps) video dimension, but high in audio dimension. The reported study categorizes the effect of playout rates on audio-visual demonstrations with different VCS schema values. This study, while important, does not cover the loss parameters, transient mis synchronizations, and rate fluctuations, all of which can happen during audio-visual display. The Likert scores of [AFKN94] is from 1 to 7, whereas our scale is from 1 to 10. In addition, we also use the scale of [Ste96]. One of the advantages of this study is the block design of the experiment, in which the combined effect of multiple parameter variations on perception were determined, whereas, in our experiment, we have only determined the effects of individual parameters.

8.4 Limitations of the current experiment and our ongoing work

The aggregate loss experiment for audio needs to be redone with appropriate clips, since we eliminated silence rather than speech. We are also in the process of comparing our results with known perceptual studies of silence elimination. Another parameter we would like to know is the perceptual difference between skipping video frames versus repeating the same frame. These are different policies, between which our current metrics do not distinguish.

Secondly, we would like to understand the combined effect of our parameters on human perception. In this respect, combining our results with those of other studies to obtain a combined Likert scale as a function with multiple inputs as defects will be most beneficial. We are also planning a block-designed factorial [Edw85] experiment involving more QoS parameters. As stated, this involves having a sufficiently randomized experiment where the participant's boredom does not affect their judgment. Some of our ongoing work addresses this issue in detail. The benefits of such a study are significant in the implementation of multimedia testbeds, as given below.

- It allows the prioritization of user needs.
- It allows for the most beneficial dynamic QoS adjustments [AFKN94].
- It adds up to building a comprehensive user-level QoS metric for multimedia [Sta96].
- It helps in resource management [Sta96].
- It helps in exception handling and fault tolerance [Nai96].
- It can be used in multimedia server design.

We are also in the process of enhancing the Tcl/Tk-based [Wel95, Ous94] Berkeley Continuous Media Toolkit (CMT) [SRY93] to enhance its performance by using our new-found tolerances to defects reported in this paper. In this work, we see a clear need for a comprehensive QoS metric.

9 Conclusions

Based on the observation that (1) loss of media content, (2) rate variations and (3) the degree of transient missynchronizations result in user discontent in multimedia presentations, we designed metrics to measure these phenomena. A user study was carried out to substantiate our initial observations, and thereby validate the assumptions that underly our model. The results of this study and its analysis have been presented. Finally, the usage of our experimental results in multimedia system design has been discussed.

References

[AFKN94] Aptekar RT, Fisher JA, Kisimov V, Nieshlos H (1994) Distributed Multimedia:User Perception and Dynamic QoS. In SPIE 2188:226–234

[Edw85] Edwards AE (1985) Experimental Design In Psychological Research, 5th edition. Harper & Row, New York

[Geo96] Georganas ND (1996) Synchronization issues in multimedia presentational and conversational applications. In Proceedings of the 1996 Pacific Workshop on Distributed Multimedia Systems (DMS'96), June 1996, Hong Kong, invited talk

[HRKHS96] Huang J, Richardson J, Kenchamanna-Hosekote DR, Srivastava J (1996) Presto: Final technical report. Technical report, Honeywell Technology Center, Minneapolis, MN

[Nai96] Naik K (1996) Exception handling and fault-tolerence in multimedia synchronization. IEEE J Sel Areas Commun 14(1):196–211

[Opp83] Oppenheim AN (1983) Questionnaire Design and Attitude Measurement. Heinemann, London

[Ous94] Ousterhout JK (1994) Tcl and the Tk Toolkit. Addison-Wesley, Reading, Mass.

[PESMI96] Part-Enander E, Sjoberg A, Melin B, Isaksson P (1996) The Matlab Handbook. Addison-Wesley, Reading, Mass.

[SNL95] Schmidt B, Northcutt J, Lam M (1995) A Method and Apparatus for Measuring Media Synchronization. In Gusella R, Little, TDC (eds) Proceedings of the 5th International Workshop on Networks and Operatiing System Support for Video and Audio (NOSDAV '95) volume 5, April 1995, Durham, N.H., pp 203–214

[SRY93] Smith B, Rowe L, Yen S (1993) A Tcl/Tk Continuous Media Player. In Proceedings of the Tcl-Tk Workshop, June 1993, Berkeley, CA

[Sta96] Staehli RA (1996) Quality of Service Specification for Resource Management in Multimedia. PhD thesis. Oregon Graduate Institute of Science and Technology, OR

[Ste96] Steinmetz R (1996) Human perception of jitter and media synchronization. IEEE J Sel Areas Commun 14(1):61–72

[SB96] Steinmetz R, Blakowski G (1996) A media synchronization survey: Reference model, specification and case studies. IEEE J Sel Areas Commun 14(1):5–35

[SGN96] Steinmetz R, Georganas ND, Nakagawa T (1996) Guest editorial: Synchronization issues in multimedia communications. IEEE J Sel Areas Commun 14(1):1–4

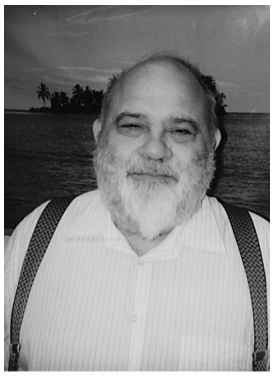
[Tow93] Towsley D (1993) Providing quality of service in packet switched networks. In Donatiello L, Nelson R (eds) Performance Evaluation of Computer Communication Systems. Springer, Berlin Heidelberg New York, pp 560–586

[Wel95] Welch B (1995) Practical Programming in Tcl and Tk. Prentice Hall, Englewood Cliffs, N.J.

[WS96] Wijesekera D, Srivastava J (1996) Quality of Service (QoS) Metrics for Continuous Media. Multimedia Tools Appl 3(1):127–166

DUMINDA WIJESEKERA is a senior engineer at the space systems division of Honeywell Inc., in Clearwater, Fla. His current work involves toolkit support for embedded high-performance computing for civilian and military domains. He has a PhD in Mathematical Logic from Cornell University and a PhD in Computer Science from the University of Minnesota, and prior to the current position has worked as an assistant professor at the University of Wisconsin and as a visiting post-doctoral fellow at the Army High-Performance Computing Research Center at the University of Minnesota. His areas of interests are quality of service in multimedia, datamining, formal methods in software engineering and high-performance computing.

JAIDEEP SRIVASTAVA received the B.Tech. degree in computer science from the Indian Institute of Technology, Kanpur, India, in 1983, and the M.S. and Ph.D. degrees in computer science from the University of California, Berkeley, in 1985 and 1988, respectively. Since 1988 he has been on the faculty of the Computer Science Department, University of Minnesota, Minneapolis, where he is currently an Associate Professor. In 1983, he was a research engineer with Uptron Digital Systems, Lucknow, India. He has published over 110 papers in refereed journals and conferences in the areas of databases, parallel processing, artificial intelligence, and multimedia. His current research is in the areas of databases, distributed systems, and multimedia computing. He has given a number of invited talks and participated in panel discussions on these topics. Dr. Srivastava is a senior member of the IEEE Computer Society and the ACM. His professional activities have included being on various program committees, and refereeing for journals, conferences, and the NSF.



ANIL NERODE is Goldwin Smith professor of mathematics and computer science at Cornell University, Ithaca, New York. He received his Ph.D. under Saunders MacLane at the University of Chicago in 1956, spent 1957–1958 at the Institute for Advanced Study in Princeton with K. Godel as a postdoctoral fellow, and 1958–1959 with Alfred Tarski at the University of California, Berkeley as a visiting assistant professor. He joined the Cornell faculty as assistant professor of mathematics at the invitation of J. Barkley Rosser in 1959, and has been there ever since. He served as Chair of the Mathematics Department (1982–1987), as Director of the Mathematical Sciences Institute (1987–1997), and is currently Director of the Center for the Foundations of Intelligent Systems. He has been an editor for many journals, including the Journal of Symbolic Logic (1968–1973), the Annals of Pure and Applied Logic (1987–1997), the Annals of Mathematics and Artificial Intelligence (1990–present), and the Journal of Pure and Applied ALgebra (1990–present). He is the author of over 100 research papers in mathematical logic, recursive functions, automata, computational complexity, distributed autonomous control, and hybrid systems. He has been co-author or co-editor of numerous books, and chair or plenary speaker in numerous conferences in mathematical logic, control engineering, and computer science. He has directed 38 Ph.D. dissertations. He has also served a term as vice-president of the American Mathematical Society.

MARK D. FORESTI is currently a computer scientist with the Air Force Research Lab located in Rome NY. He received his BS and MS degrees in Computer Science from the Syracuse University and The State University of New York in 1985 and 1991, respectively. His research interests are in various areas of information technologies, including Internet information management, multimedia, rapid application development, and distributed collaboration. Over the past several years Mr. Foresti has been supporting DARPA in developing new capabilities and technologies for transition to support Air Force information management requirements.