PAPER

Media Synchronization Quality of Reactive Control Schemes**

Yutaka ISHIBASHI[†a)]**,** ***Regular Member,*** **Shuji TASAKA**[†]**,** ***Fellow, and*** **Hiroki OGAWA**[††*]**,** ***Nonmember***

SUMMARY This paper assesses the media synchronization quality of recovery control schemes from asynchrony, which are referred to as reactive control schemes here, in terms of objective and subjective measures. We deal with four reactive control techniques: skipping, discarding, shortening and extension of output duration, and virtual-time contraction and expansion. We have carried out subjective and objective assessment of the media synchronization quality of nine schemes which consist of combinations of the four techniques. The paper makes a comparison of media synchronization quality among the schemes. It also clarifies the relations between the two kinds of quality measures.

key words: *media synchronization, reactive control, quality comparison, objective assessment, subjective assessment*

1. Introduction

Media synchronization control is one of key techniques for realizing distributed multimedia applications such as multimedia conferencing and distance learning over the Internet. The control is necessary for preserving the timing relation between *media units* (*MUs*) such as video frames in a single stream and the temporal relation among multiple media streams by compensating for network delay jitter [1], [2]. The former is referred to as *intra-stream* synchronization control, and the latter is called *inter-stream* synchronization control. Lip synchronization is a typical example of inter-stream synchronization, and it means synchronization between spoken voice and the movement of the speaker's lips (i.e., video). For high quality of lip synchronization, we need both types of synchronization control.

A number of media synchronization algorithms have been proposed to meet diverse requirements [3]–[11]. However, the relationships (especially, the quantitative relations) among the algorithms are not sufficiently clear. One of the reasons is that the situations and backgrounds in which algorithms have been proposed are different from each other. Thus, it is difficult to make a performance comparison among the algorithms on the same conditions. Also, there is no widely-used performance measure for media synchronization algorithms.

In order to clarify the relationships, the authors made a survey of media synchronization algorithms [12][***]. Then, they grouped media synchronization control techniques used in the algorithms into four categories: *basic control*, *preventive control*, *reactive control*, and *common control*. The basic control techniques are needed in almost all the algorithms, and they are indispensable to preserve the temporal relationships among media streams. The preventive control techniques are required to try to avoid asynchrony (i.e., out of synchronization). Thus, the techniques are used before asynchrony occurs. The reactive control techniques are employed to recover from asynchrony after it has occurred. The common control techniques can be used as both preventive and reactive control ones. In each category, the techniques are further classified into two groups by locations at which they are employed (namely, media sources or destinations). In [12], however, there is no quantitative discussion.

This paper assesses the media synchronization quality of reactive control schemes employed at destinations in terms of subjective and objective measures. The main purpose of this paper is to establish an assessment method of media synchronization quality of reactive control schemes. Therefore, we focus mainly on the assessment method here. We consider reactive control schemes which are employed in a number of algorithms. We pick up the following four techniques from the reactive control techniques used at destinations: *skipping*, *discarding*, *shortening and extension of output duration*, and *virtual-time contraction and expansion* [12]. We have carried out subjective and objective assessment of the lip synchronization quality of nine schemes which consist of combinations of the four techniques. The paper makes a comparison of the quality among the schemes. It also investigates the relations between the two kinds of quality measures in order to clarify what kinds of objective measures are more important than the others as in [13] and [14].

Manuscript received December 18, 2002.
Manuscript revised April 21, 2003.

[†]The authors are with the Department of Computer Science and Engineering, Graduate School of Engineering, Nagoya Institute of Technology, Nagoya-shi, 466–8555 Japan.

[††]The author was with the Department of Electrical and Computer Engineering, Nagoya Institute of Technology, Nagoya-shi, 466–8555 Japan.

[*]Presently, with NTT DATA Corp.

a) E-mail: ishibasi@nitech.ac.jp

**This paper was presented in part at IEEE INFOCOM 2001, Anchorage, AK, USA, April 2001.

***Note that as described in [12], we take a step-by-step approach in order to clarify the relationships among the algorithms. We first made a survey of the algorithms. Next, we picked up all the synchronization control techniques employed in the algorithms and classified the techniques into several categories. Then, we evaluate the effect of each technique on the quality on the same conditions; we are currently at this stage. Finally, we can make a quantitative comparison of the quality of various combinations of the techniques; at this time, we plan to treat the algorithms directly.

The rest of this paper is organized as follows. In Sects. 2 and 3, we describe the reactive control techniques and the nine schemes, respectively. We also explain the method of subjective assessment and that of objective assessment in Sect. 4. Furthermore, assessment results are presented in Sect. 5. Section 6 concludes the paper.

2. Reactive Control Techniques

Here we suppose that voice and video streams input at a source are transferred to a destination via a network like the Internet.

In [12], the reactive control techniques used at the destination are grouped into the following four types.

(1) Skipping (discarding) and pausing (repeating)†
When the output timing of the current MU is late, the destination skips the succeeding MUs if the MUs arrives earlier [4], [5], [7]–[10], [15]–[19]. We can also discard late MUs [11], [20]; however, discarding late MUs may lead to poor average MU rate, which is defined as the average number of MUs output per second at the destination.
For MPEG video, the condition of picture skipping is changed depending on the picture type [19], [21], [22].
When the buffer starvation occurs, the destination pauses output of MUs. This means that for video the destination continues outputting the previous MU until the next MU becomes available. It is also possible to output other data at this time.

(2) Shortening and extension of output duration
In order to recover from asynchrony gradually without serious degradation of the output quality (that is, so as not to be noticed by users), the destination shortens or extends the output duration of each MU until the recovery from asynchrony [7]–[9], [17], [23], [24].
Shortening of the output duration of MUs includes fast-forwarding (without skipping) of MUs, and extension leads to pausing output of MUs.

(3) Virtual-time contraction and expansion
In addition to the actual time, we introduce a virtual-time which expands or contracts according to the amount of delay jitter of MUs received at the destination, and media are rendered along the virtual-time axis. In [8] and [9], the virtual-time contraction and expansion are realized in a form of modification of the *target output time*, which is the time when the destination should output an MU††. In [25], the *slide control* scheme, which changes the amount of the modification of the target output time according to the extent of asynchrony, is proposed and applied to PHS (Personal Handy Phone System) [26]. In [4], the Logical Time System (LTS) corresponds to the virtual-time. Only virtual-time expansion is performed by stopping the LTS temporarily. Virtual-time expansion is also exploited in [27] and [28]. The set-back and advance operations of the PlayOut Clock in [10] correspond to virtual-time expansion and contraction, respectively.
This technique differs from shortening and extension of output duration in that the former indirectly changes the output timing by modifying the virtual-time (equivalently, resetting the origin of the time axis), while the latter directly does (that is, the origin of the time axis is kept the same).

(4) Master-slave switching
For inter-stream synchronization control, the roles of the *master* and *slave* streams††† can be changed dynamically [7], [10]. When the amount of asynchrony for a slave stream becomes large, the destination switches the stream from slave to master and performs the appropriate adaptation.

We focus on the skipping, discarding, shortening and extension of output duration, which is here called the *SE control* for short, and virtual-time contraction and expansion (referred to as the *VT control*) in this paper. We also employ the pausing control together with each of the four types of control. This is because pausing occurs if there is no MU to be output at the destination. We assume in this paper that when pausing occurs, the destination continues outputting the previous MU for video until the next MU becomes available, and pausing for voice leads to a silence.

As described earlier, we handle lip synchronization in this paper, where the voice stream is selected as the master (denoted by stream 1), and the video stream as the slave (stream 2) [4], [8]. This is the reason why we do not deal with the master-slave switching control here.

As common control techniques which can be used for reactive control at the destination, we have the following two.

(5) Adjustment of output rate
The clock frequency of output device (i.e., hardware output rate) is adjusted according to the synchronization quality [4]. In [7], [18], and [29], the destination adjusts the output rate depending on the length of the queue waiting for output††††.

(6) Interpolation of data
We can interpolate data so as to adjust the effective output rate [4].

---

†In [12], these techniques are referred to as reactive skipping (discarding) and reactive pausing (repeating) so that we can distinguish the techniques from preventive skipping (discarding) and preventive pausing (repeating), respectively.

††If there were no network delay jitter, it would be the arrival time of the MU, which is equal to its departure time at the source plus the network propagation delay. In reality, however, the jitter exists. Therefore, the target output time is modified in order to adapt to the delay variation.

†††We can categorize the streams into a master stream and slave streams, which are synchronized with the master stream. Generally, a media stream which is most sensitive to intra-stream synchronization error is selected as the master, and the others as slaves.

††††In [18], Yuang et al. employ the intelligent video smoother (IVS) [30], which is a neural network traffic predictor, for intra-stream synchronization of video.

However, we do not treat these two in this paper since they need special hardware or software implementation.

It is possible to employ two or more reactive control techniques together for each stream. We can also use a reactive control technique for the voice and another one for the video. Furthermore, the reactive control techniques can be utilized for inter-stream synchronization as well as intra-stream synchronization. In this paper, we assume that intra-stream synchronization control is carried out over the master stream [8], and that only inter-stream synchronization control is exerted over the slave stream for simplicity.

3. Reactive Control Schemes

This paper deals with the following nine reactive control schemes, which are different from each other in reactive control techniques used for the master and slave streams:

(1) the *discarding/discarding* scheme (this notation represents a technique used for the master/that for the slave),
(2) the *skipping/skipping* scheme,
(3) the *SE/SE* scheme,
(4) the *skipping+SE/skipping+SE* scheme ('+' means that multiple techniques are used together),
(5) the *skipping+VT/skipping* scheme,
(6) the *SE+VT/SE* scheme,
(7) the *SE+VT/skipping* scheme,
(8) the *SE+VT/skipping+SE* scheme, and
(9) the *skipping+SE+VT/skipping+SE* scheme.

As described later in this section (in Sect. 3.1 4)), the VT control technique is used for the master stream together with other techniques which determine the output time of each MU since the VT control treats only a virtual-time axis. We also handled the other combinations of the techniques. However, we do not handle the combinations in this paper since we confirmed that they each have almost the same results as some of the above schemes in our preliminary objective assessment.

As basic control techniques, each of the nine schemes employs the following two: *attachment of timestamps to MUs* and *buffering of MUs* [12].

Media streams are categorized into live and stored media streams. In this paper, we assume that live media have more severe constraints on the time intervals from generation time to output time than stored media. We here handle only live media. For stored media case, the reader is referred to [31].

In what follows, we explain how to determine the output time of each MU under the intra-stream and inter-stream synchronization control for live media. As the first step of our research, we here adopt the determination methods for simplicity.

3.1 Intra-Stream Synchronization Control

In the intra-stream synchronization control, the destination calculates the *ideal target output time* [32] of each MU, which is defined as the time at which the MU should be output in the case of no network delay jitter. Then, the destination determines the output time of the MU by using the ideal target output time.

• **Case of** $m = 1$
$D_1^{(1)} = A_1^{(1)} + J_{\max}$
IF $D_1^{(1)} - T_1^{(1)} < \Delta_{\rm al}$
THEN $x_1^{(1)} = D_1^{(1)}$
ELSE $x_1^{(1)} = T_1^{(1)} + \Delta_{\rm al}$
• **Case of** $m \geqq 2$
$x_m^{(1)} = x_1^{(1)} + T_m^{(1)} - T_1^{(1)}$

Fig. 1 Calculation of the ideal target output time for live media.

For the description of the schemes, we denote the ideal target output time of the m-th MU ($m = 1, 2, \cdots$) in stream 1 (i.e., the master stream) by $x_m^{(1)}$. Also, let $J_{\max}$ represent an estimate of the maximum network delay jitter (i.e., the buffering time of the first MU [8]). Let $T_m^{(j)}$, $A_m^{(j)}$, and $D_m^{(j)}$ denote the generation time (i.e., the timestamp), arrival time, and output time, respectively, of the m-th MU in stream j ($j = 1$ or 2).

We show how to calculate $x_m^{(1)}$ for live media in Fig. 1 (see [31] for stored media). The output time $D_1^{(1)}$ of the first MU ($m = 1$) is set to the arrival time $A_1^{(1)}$ plus $J_{\max}$. The destination determines $x_1^{(1)}$ by using $D_1^{(1)}$ and then calculates $x_m^{(1)}$ of the m-th MU ($m = 2, 3, \cdots$).

For live media, the real-time property is very important. Therefore, if the MU delay $D_1^{(1)} - T_1^{(1)}$ is larger than the maximum allowable delay $\Delta_{\rm al}$ [9], we set $x_1^{(1)} = T_1^{(1)} + \Delta_{\rm al}$[†] as shown in Fig. 1. Otherwise, we set $x_1^{(1)} = D_1^{(1)}$.

1) Discarding and skipping control

Figures 2 and 3 illustrate how the output time $D_m^{(j)}$ of the m-th MU ($m \geqq 2$) in stream j is determined under the discarding control and skipping control, respectively. Note that we here explain the determination method of the output time for stream j since the two kinds of technique are used in the intra-stream and inter-stream synchronization control.

In Figs. 2 and 3, if the m-th MU in stream j arrives at the destination by its ideal target output time $x_m^{(j)}$, the destination sets $D_m^{(j)} = x_m^{(j)}$ in the two types of control. Otherwise, the discarding control discards the m-th MU (see Fig. 4). Under the skipping control, the destination skips the m-th MU if the $(m + 1)$-st MU has already arrived and if the m-th MU has arrived more than a constant time late (see

[†]Globally synchronized clocks [1], [9] are assumed to be used in this paper. Using Network Time Protocol (NTP) [33] over the Internet, we can adjust the clock ticks to each other within a few milliseconds. This resolution is high enough for lip synchronization.

IF $A_m^{(j)} \leqq x_m^{(j)}$
THEN $D_m^{(j)} = x_m^{(j)}$
ELSE discard the m-th MU

Fig. 2 Determination of the output time under the discarding control.

IF $A_m^{(j)} \leqq x_m^{(j)}$
THEN $D_m^{(j)} = x_m^{(j)}$
ELSE IF the $(m+1)$-th MU has arrived **AND** $A_m^{(j)} \geqq x_m^{(j)} + T_{m+1}^{(j)} - T_m^{(j)}$
THEN skip the m-th MU
ELSE $D_m^{(j)} = A_m^{(j)}$

Fig. 3 Determination of the output time under the skipping control.

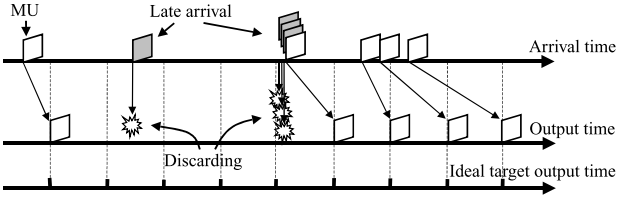

Fig. 4 An example of the output timing of MUs under the discarding control.

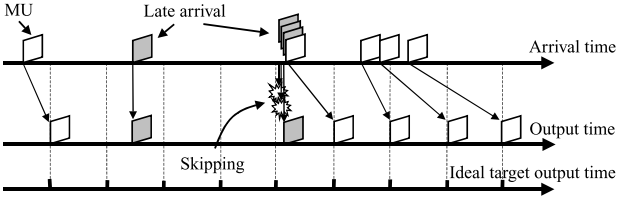

Fig. 5 An example of the output timing of MUs under the skipping control.

Fig. 5). For simplicity, in this paper, we set the constant time to the difference in timestamp between the $(m+1)$-st and m-th MUs (i.e., $T_{m+1}^{(j)} - T_m^{(j)}$)† Otherwise, the destination outputs the m-th MU on receiving it.

2) SE control

We display the determination method of the output time of the m-th MU in stream j under the SE control in Fig. 6. If the m-th MU arrives neither by $x_m^{(j)}$ nor by $D_{m-1}^{(j)} + T_m^{(j)} - T_{m-1}^{(j)}$, the destination sets $D_m^{(j)} = A_m^{(j)}$. Otherwise, when $x_m^{(j)} \leqq D_{m-1}^{(j)} + T_m^{(j)} - T_{m-1}^{(j)}$, the destination tries to reduce the output duration between the $(m-1)$-st and m-th MUs by a constant time c in order to make the output time approach the ideal target output time gradually (see Fig. 7). However, decreasing the output duration by c may

IF $A_m^{(j)} \leqq \max(x_m^{(j)}, D_{m-1}^{(j)} + T_m^{(j)} - T_{m-1}^{(j)})$
THEN IF $x_m^{(j)} \leqq D_{m-1}^{(j)} + T_m^{(j)} - T_{m-1}^{(j)}$
THEN $D_m^{(j)} = \max(D_{m-1}^{(j)} + T_m^{(j)} - T_{m-1}^{(j)} - c, x_m^{(j)}, A_m^{(j)})$
ELSE $D_m^{(j)} = \min(\max(D_{m-1}^{(j)} + T_m^{(j)} - T_{m-1}^{(j)} + c, A_m^{(1)}), x_m^{(j)})$
ELSE $D_m^{(j)} = A_m^{(j)}$

Fig. 6 Determination of the output time under the SE control.

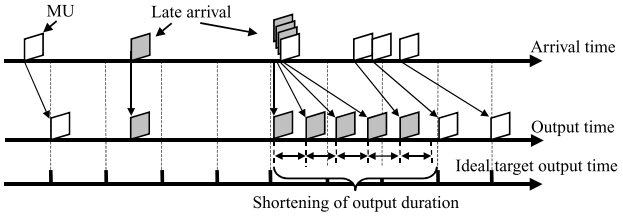

Fig. 7 An example of the output timing of MUs under the SE control.

produce the output time earlier than $x_m^{(j)}$ or $A_m^{(j)}$. To avoid this, the destination selects the latest time from among $x_m^{(j)}$, $A_m^{(j)}$, and $D_{m-1}^{(j)} + T_m^{(j)} - T_{m-1}^{(j)} - c$, which is the time instant obtained by reducing the output duration by c. When $x_m^{(j)} > D_{m-1}^{(j)} + T_m^{(j)} - T_{m-1}^{(j)}$, the destination extends the output interval between the $(m-1)$-st and m-th MUs by c in order to make the output time approach $x_m^{(j)}$††. In this case, it is also possible to try to output the m-th MU earlier than $A_m^{(j)}$ or later than $x_m^{(j)}$. Therefore, the destination compares $D_{m-1}^{(j)} + T_m^{(j)} - T_{m-1}^{(j)} + c$ (that is, the time instant obtained by increasing the output duration by c) with $A_m^{(j)}$, and chooses the later time, which is also compared with $x_m^{(j)}$. Then, it selects the earlier as $D_m^{(j)}$.

3) Skipping+SE control

In the skipping+SE control, if an MU arrives largely late under the SE control, the destination tries to skip the MU. We show the determination method of $D_m^{(j)}$ under the skipping+SE control in Fig. 8. The method can be obtained by replacing the bottom line in Fig. 6 with the third through the bottom lines in Fig. 3. In these four lines, we employ $D_{m-1}^{(j)} + T_m^{(j)} - T_{m-1}^{(j)}$ instead of $x_m^{(j)}$.

4) VT control

The VT control technique is applied to only the master

†In [9], the time is set to 160 ms. As a result of the quality assessment with this value, we obtained almost the same results as those in this paper, where the value is set to 50 ms (see Table 1).

††This extension can occur under the inter-stream synchronization control in Sect. 3.2.

IF $A_m^{(j)} \leqq \max(x_m^{(j)}, D_{m-1}^{(j)} + T_m^{(j)} - T_{m-1}^{(j)})$
  THEN IF $x_m^{(j)} \leqq D_{m-1}^{(j)} + T_m^{(j)} - T_{m-1}^{(j)}$
    THEN $D_m^{(j)} = \max(D_{m-1}^{(j)} + T_m^{(j)} - T_{m-1}^{(j)} - c, x_m^{(j)}, A_m^{(j)})$
    ELSE $D_m^{(j)} = \min(\max(D_{m-1}^{(j)} + T_m^{(j)} - T_{m-1}^{(j)} + c, A_m^{(1)}), x_m^{(j)})$
  ELSE IF the $(m+1)$-th MU has arrived **AND** $A_m^{(j)} \geqq D_{m-1}^{(j)} + T_{m+1}^{(j)} - T_{m-1}^{(j)}$
    THEN skip the m-th MU
    ELSE $D_m^{(j)} = A_m^{(j)}$

Fig. 8 Determination of the output time under the skipping+SE control.

stream (i.e., stream 1) [8], [9] differently from the other control techniques.

To describe the control, we define the *target output time* $t_m^{(1)}$ of the m-th MU in stream 1 as the time at which the MU should be output in the case where there exists network delay jitter [8]. Let $t_m^{(1)*}$ and $\Delta S_m^{(1)}$ denote the *modified target output time* and the *slide time* (i.e., the amount of modification) [32], respectively. Then, $t_m^{(1)}$ and $t_m^{(1)*}$ are given by

$$t_1^{(1)} = x_1^{(1)}, \tag{1}$$

$$t_m^{(1)} = x_m^{(1)} + \sum_{i=1}^{m-1} \Delta S_i^{(1)} \quad (m \geqq 2), \tag{2}$$

$$t_m^{(1)*} = t_m^{(1)} + \Delta S_m^{(1)} \quad (m \geqq 1), \tag{3}$$

where $\Delta S_1^{(1)} = 0$.

We show the determination method of $D_m^{(1)}$ $(m \geqq 2)$ under the VT control for live media in Fig. 9. In the figure, when $A_m^{(1)} \leqq t_m^{(1)}$, if $t_m^{(1)} - T_m^{(1)} > \Delta_{\rm al}$ or if each MU arrives before its target output time during a period of $T_{\rm nolate}$ until $A_m^{(1)}$, then the destination tries to advance the target output time of each of the succeeding MUs by a constant time r (see Fig. 10). However, this may make the target output time earlier than the ideal one. To avoid this, we set $\Delta S_m^{(1)} = -\min(t_m^{(1)} - x_m^{(1)}, r)$. Otherwise, we set $\Delta S_m^{(1)} = 0$. After setting $\Delta S_m^{(1)}$, the destination calculates $t_m^{(1)*}$. Then, it determines $D_m^{(1)}$ by carrying out the skipping, SE, or skipping+SE control with $t_m^{(1)*}$ instead of $x_m^{(1)}$.

When $A_m^{(1)} > t_m^{(1)}$, the destination exerts the skipping, SE, or skipping+SE control by using $t_m^{(1)}$ instead of $x_m^{(1)}$ and determines $D_m^{(1)}$. Then, if $D_m^{(1)} - t_m^{(1)}$ is larger than $T_{\rm h2}^{(1)}$, which is a threshold for the modification of the target output time for stream 1 [8], we set $\Delta S_m^{(1)} = D_m^{(1)} - t_m^{(1)}$ in order to delay the output timing of the succeeding MUs (see Fig. 10). This means the increase of the buffering time.

Thus, the VT control expands or contracts the virtual-time by changing the origin of the time axis used to determine the output time of each MU. On the other hand, the SE control does not change the time origin.

IF $A_m^{(1)} \leqq t_m^{(1)}$
  THEN IF $t_m^{(1)} - T_m^{(1)} > \Delta_{\rm al}$ **OR** each MU arrives before its target output time during a period of $T_{\rm nolate}$
    THEN $\Delta S_m^{(1)} = -\min(t_m^{(1)} - x_m^{(1)}, r)$
    ELSE $\Delta S_m^{(1)} = 0$
    $t_m^{(1)*} = t_m^{(1)} + \Delta S_m^{(1)}$
    carry out the skipping, SE, or skipping+SE control by using $t_m^{(1)*}$ instead of $x_m^{(1)}$, and determine $D_m^{(1)}$
  ELSE carry out the skipping, SE, or skipping+SE control by using $t_m^{(1)}$ instead of $x_m^{(1)}$, and determine $D_m^{(1)}$
    IF $D_m^{(1)} - t_m^{(1)} > T_{\rm h2}^{(1)}$
      THEN $\Delta S_m^{(1)} = D_m^{(1)} - t_m^{(1)}$
      ELSE $\Delta S_m^{(1)} = 0$

Fig. 9 Determination of the output time under the VT control for live media.

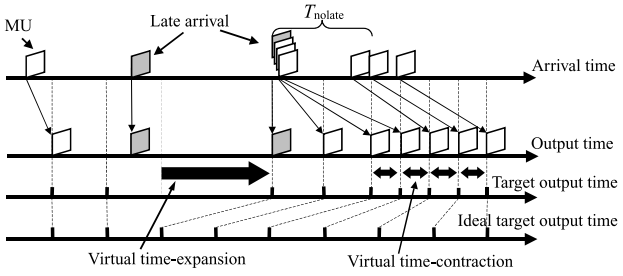

Fig. 10 An example of the output timing of MUs under the VT control.

3.2 Inter-Stream Synchronization Control

In this control, the output timing of the slave stream is determined according to that of the master one. Figure 11 shows how to determine the output time $D_m^{(2)}$ of the m-th MU in the slave (stream 2). We here assume that the master and slave streams are *loosely coupled* [8]. In this case, on the conditions that the latest output MU in stream 1 is the n-th one when the m-th MU in stream 2 arrives at the destination, the output time $D_m^{(2)}$ of the m-th MU in stream 2 is calculated by using the *derived output time* [8], which denotes the output time of the corresponding master MU (the n-th MU) plus the relative generation time of the slave MU to the master MU (i.e., $D_n^{(1)} + T_m^{(2)} - T_n^{(1)}$) (see Fig. 12).

In Fig. 11, for simplicity, we assume that $T_1^{(2)} \geqq T_1^{(1)}$ for the first slave MU ($m = 1$). Thus, the destination waits for the output of the first master MU if the MU has not been output yet. Otherwise, the destination compares $D_1^{(1)} + T_1^{(2)} - T_1^{(1)}$ and $A_1^{(2)}$ and then selects the later time as $D_1^{(2)}$. In the case of $m \geqq 2$, it determines $D_m^{(2)}$ by using $D_n^{(1)} + T_1^{(2)} - T_n^{(1)}$ instead of $x_m^{(1)}$ under the discarding,

• **Case of** $m = 1$

IF the first master MU has not output yet

THEN wait for output of the first master

ELSE $D_1^{(2)} = \max(D_1^{(1)} + T_1^{(2)} - T_1^{(1)}, A_1^{(2)})$

• **Case of** $m \geqq 2$

carry out the skipping, discarding, SE, or skipping+SE by using $D_n^{(1)} + T_m^{(2)} - T_n^{(1)}$ instead of $x_m^{(1)}$, and determine $D_m^{(2)}$

Fig. 11 Determination of the output time under the inter-stream synchronization control.

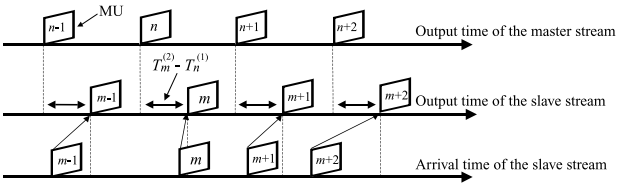

Fig. 12 An example of the output timing of MUs under the inter-stream synchronization control.

skipping, SE, or skipping+SE control.

4. Methodology for Quality Assessment

In order to make a comparison of the lip synchronization quality among the nine schemes, we have assessed the quality subjectively and objectively. We here describe the assessment method.

4.1 Subjective Assessment

We employ computer simulation to perform the lip synchronization control in this paper. The purpose of using computer simulation is to get rid of the influence of the difference in implementation among the nine schemes on the lip synchronization quality. By outputting the voice and video streams actually (see Fig. 13) and Table 1[†] with the output timing of MUs obtained in the simulation, we have carried out subjective assessment of the quality.

We have used a workstation (Sun Ultra 2; Solaris 2.5.1) for simulation and for output of the voice and video streams, which are stored at the workstation together with timestamps. We use the stored voice and video streams even for the assessment of the live media synchronization quality in order to output the voice and video streams with the same contents in each test sample[††].

In the simulation, we generate a pseudo-network delay which is normally distributed with a mean of 100 ms for each MU [7][†††]; the standard deviation is varied. When a generated value is negative, we reset the value to zero[††††]. Since the main purpose of this paper is to establish an assessment method of media synchronization quality of reactive control schemes as described in Sect. 1, for simplicity, we employ this model for simulating network delays. In our

Fig. 13 A sample image of video.

Table 1 Specifications of voice and video.

item	voice	video
coding scheme	ITU-T G.711 μ-law	JPEG
image size [pixels]	-	320×240
average MU rate [MU/s]	20	
average bit rate [kb/s]	64	1047

assessment method, we can easily employ a variety of models (including trace data obtained from the Internet or a network simulator). To demonstrate this, we adopt the model as an example in this paper. What kinds of models we should employ is another important issue to be addressed, which is outside the scope of this paper; this is for further study.

Also, as the generation time of each MU, we adopt the timestamp of the MU in the simulation. We set $T_{\mathrm{h2}}^{(1)} = 320$ ms, $c = r = 20$ ms [9], $J_{\max} = 100$ ms, $\Delta_{\mathrm{al}} = 400$ ms, and $T_{\mathrm{nolate}} = 5$ seconds. Furthermore, in order to avoid the disturbance of output timing of the voice and video owing to processing overhead at the workstation, we set the minimum output duration of each voice MU to 1 ms and that of each video MU to 10 ms. The minimum output duration denotes the minimum time from the moment the destination has output an MU until the instant it becomes possible to output the next MU. These values are set to be larger than the maximum processing time of each MU, which was measured in a preliminary experiment.

For the subjective assessment, we enhance the single

[†]In this paper, as a distributed multimedia application, we suppose a multimedia conferencing system which multiple users attend. In the system, several images of users' head views are displayed at each destination. Therefore, the display size is set to 320×240 pixels here.

[††]When we deal with live media, the interactivity as well as the media synchronization quality is important. For simplicity, however, we focus only on the media synchronization quality in this paper.

[†††]We also carried out the same assessment with the exponential distribution. We obtained similar results to those in this paper.

[††††]We assume in this paper that the destination carries out MU sequencing by using timestamps (or the sequence numbers). Since a negative pseudo-network delay is reset to zero, an MU with the pseudo-network delay of zero may arrive at the application layer of the destination at the same time as the previous MU owing to the MU sequencing (we here suppose that media synchronization control is carried out in the application layer). Actually, we can sometimes observe simultaneous arrival in real networks.

stimulus method in ITU-R BT.500-5 [34], which is a recommendation for subjective assessment of television pictures, as in [13] and [14]. From a small set of preliminary tests, we deduced that the value of 25 seconds of a test sample is sufficiently long for getting the opinions of assessors. In each session, which lasts for approximately 40 minutes, an assessor is presented with a series of test samples. The test samples, which are made by various combinations of the standard deviation of network delay and each scheme, were presented in random order. The number of assessors is 16. They are non-expert in the sense that they are not directly concerned with voice and video quality as part of their normal work. Their ages are between 22 and 24.

At the beginning of each session, an explanation is given to an assessor about the type of assessment and the grading scale shown in Table 2. Assessors are shown examples. They are asked to base their judgments on the overall impression given by each test sample, and to express the judgments in terms of the wording used to define the subjective scale (Table 2). Each assessor gives a score from 1 through 5 to each test sample. Then, we gather all the scores and check their coherence by calculating the mean values and the standard deviations according to the method in [34]. Thus, we express the subjective quality of the test samples by *mean opinion score* (*MOS*) [34].

4.2 Objective Assessment

For the objective assessment, we have measured the performance in each test of the subjective assessment in terms of the following measures: the *average MU rate*, *total pause time*, *average MU delay*, and *mean square error of inter-stream synchronization* [9], [13], [14], [21], [22], [25], [26], [32].

The average MU rate is defined as the average number of MUs output in a second at the destination. The total pause time denotes the sum of pausing intervals, and the interval for the m-th MU in stream j is calculated by $D_m^{(j)} - \max(t_{k+1}^{(j)}, D_k^{(j)} + T_{k+1}^{(j)} - T_k^{(j)})$ under the condition that the last MU output before the m-th one is the k-th; note that if no skipping occurs, we have $k = m - 1$. The average MU delay is the average time in seconds from the moment an MU is generated until the instant the MU is output (the MU delay of the m-th MU in stream j is $D_m^{(j)} - T_m^{(j)}$). The mean square error of inter-stream synchronization is defined as the average square of the difference between the output time of each slave MU and its derived output time (i.e., the output time of the corresponding master MU (say the n-th MU) $D_n^{(1)}$ plus the relative generation time of the slave MU (say the m-th MU) to the master $T_m^{(2)} - T_n^{(1)}$ as defined in Sect. 3.2).

Table 2 Five-grade impairment scale.

grade	impairment
5	imperceptible
4	perceptible, but not annoying
3	slightly annoying
2	annoying
1	very annoying

5. Assessment Results

We show the MOS values of the nine schemes for live media as a function of the standard deviation of network delay† in Fig. 14 (see [31] for the results in the case of stored media). We also plot the average MU rates, the total pause times, and the average MU delays for live voice and video in Figs. 15 through 20. Figure 21 illustrates the mean square error of inter-stream synchronization for live media. In all the figures, we plot the 95% confidence intervals of quality measures. However, when the interval is smaller than the size of the corresponding symbol representing the experimental result, we do not show it in the figures. Also, in the figures for voice, we do not depict the assessment results of the nine schemes but those of seven control techniques employed for voice. This is because the voice, which is the master, is not affected by the difference in control over the slave (video).

5.1 Subjective Assessment

In Fig. 14, we see that the discarding/discarding scheme has the smallest MOS value among the nine schemes when the standard deviation of network delay is larger than around 50 ms. Also, the MOS values of the four schemes which exert the skipping control over voice deteriorate largely as the standard deviation increases. This is because the output of voice sometimes breaks owing to discarding or skipping of voice MUs. Therefore, performing the discarding or skipping control over voice is not preferable.

Furthermore, from the figure, we notice that the SE+VT/SE scheme has the largest MOS value when the standard deviation is greater than around 150 ms. In this area, as the standard deviation becomes larger,

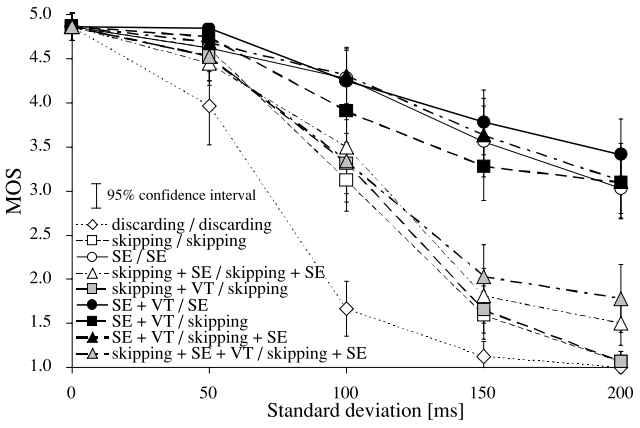

Fig. 14 MOS versus standard deviation of network delay for live media.

†When the standard deviation is 0 ms, there is no difference in output timing among all the schemes. Therefore, all the schemes have the same results in this case.

the difference in MOS value between the SE/SE and SE+VT/SE schemes increases. In addition, when the standard deviation is larger than approximately 100 ms, the SE+VT/skipping+SE scheme tends to have larger MOS values than the SE+VT/skipping scheme. The SE/SE scheme has the largest MOS value among the schemes without the VT control.

5.2 Objective Assessment

In Figs. 15 and 16, we see that when the standard deviation of network delay is larger than approximately 50 ms, the discarding scheme has the smallest average MU rate. This is because the scheme discards MUs which have arrived late. Also, in the figures, as the standard deviation becomes larger, the average MU rates of the four schemes which carry out the skipping control over the voice decrease. At the standard deviation of around 200 ms, each scheme with the VT control has a larger average MU rate than the corresponding scheme without the control. In Fig. 15, the average MU rates of the schemes which performs the SE and SE+VT control over the voice do not decrease even if the standard deviation increases. This means that all the voice MUs are output. From Fig. 16, we find that for the standard deviations larger than around 100 ms, the SE+VT/SE and the SE+VT/skipping+SE schemes have the largest and the second largest average MU rates, respectively, among the three schemes which exert the SE+VT control over the voice.

We notice in Figs. 17 and 18 that the discarding/discarding scheme has the largest total pause time when the standard deviation is larger than around 50 ms. In the case where the SE+VT control is exerted over the voice, the total pause time at the standard deviation of approximately 200 ms is smaller than that at around 150 ms. The reason is that pausing longer than the threshold $T_{\mathrm{h2}}^{(1)}$ occurs more frequently at the standard deviation of around 200 ms; then, the virtual-time is largely expanded; this means that the buffering time is increased. Note that longer buffering time can compensate for larger delay jitter.

Figures 19 and 20 reveal that the average MU delay of the discarding/discarding scheme is constant independently of the standard deviation. This is because the scheme does not output MUs which have arrived late. In the figures, when the standard deviation is larger than approximately 100 ms, the schemes which performs the SE and SE+VT control over the voice have larger average MU delays than the other schemes. In the same range, each scheme with the VT control has larger average MU delays than the corresponding scheme without the control.

From Fig. 21, we observe that all the schemes excluding the SE/SE and SE+VT/SE schemes have the mean

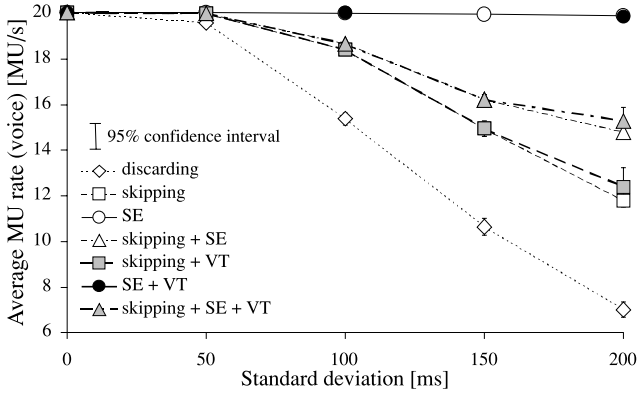

Fig. 15 Average MU rate of live voice.

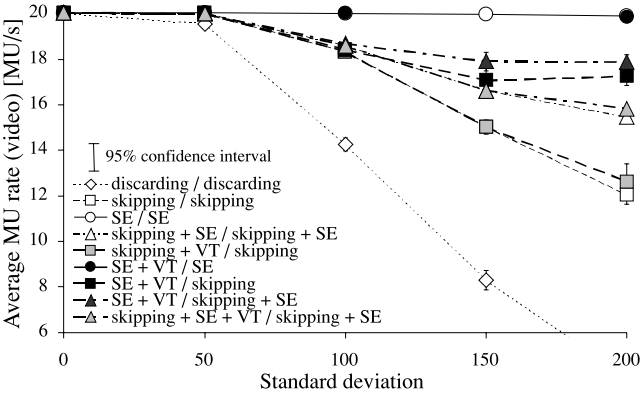

Fig. 16 Average MU rate of live video.

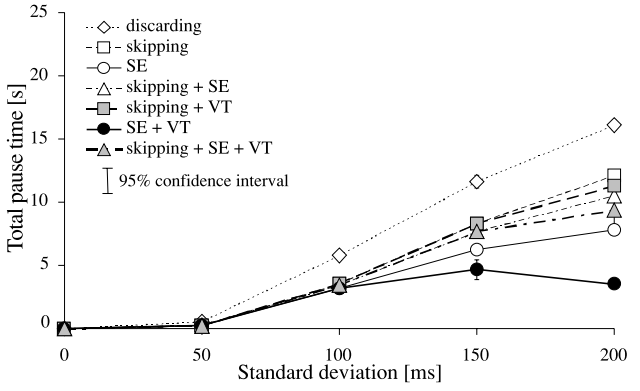

Fig. 17 Total pause time of live voice.

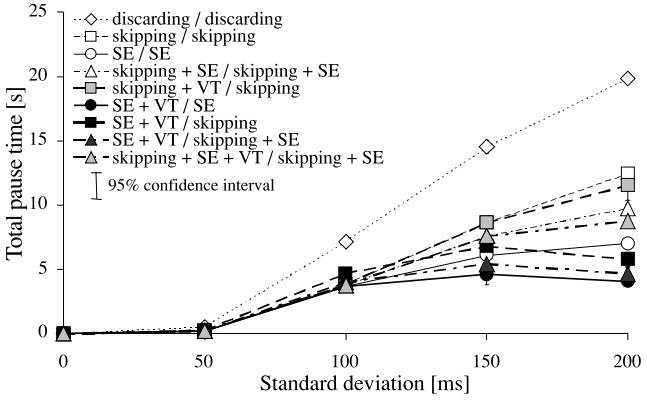

Fig. 18 Total pause time of live video.

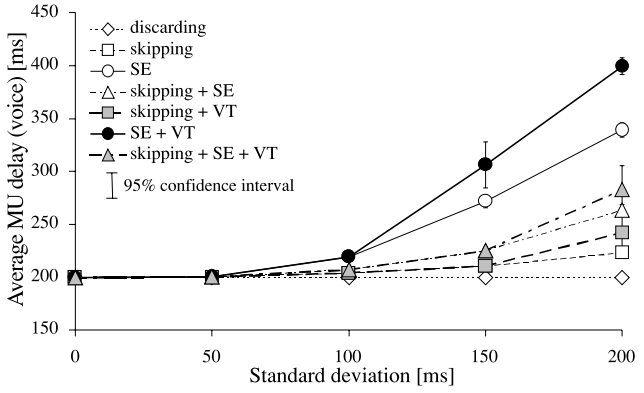

Fig. 19 Average MU delay of live voice.

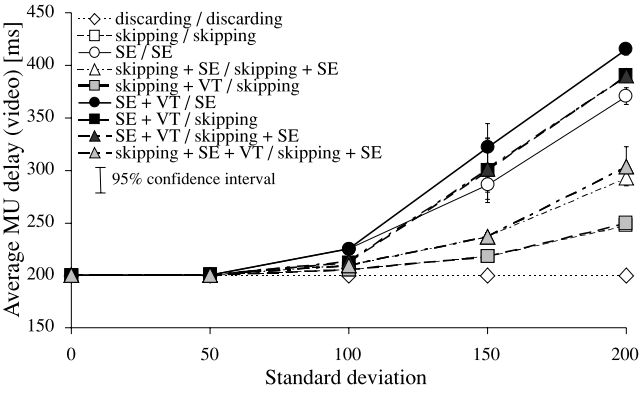

Fig. 20 Average MU delay of live video.

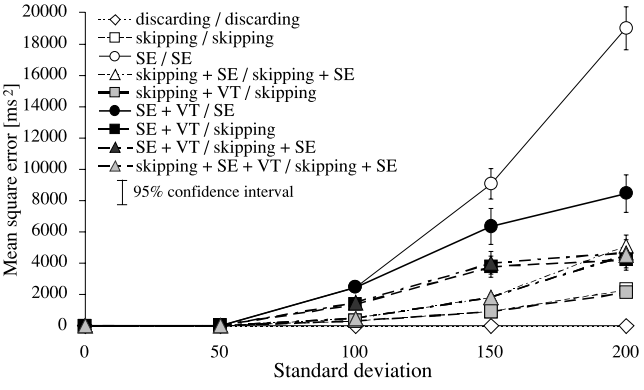

Fig. 21 Mean square error of inter-stream synchronization for live media.

square errors of inter-stream synchronization less than $6400\,\mathrm{ms}^2$ (this value is the square of 80 ms. A time difference of between -80 ms and $+80$ ms leads to lip synchronization of high quality according to [35]†) in the whole range of the standard deviation considered here. Therefore, the inter-stream synchronization quality of the schemes are very high. We also see in Fig. 21 that when the standard deviation is larger than around 150 ms, the SE/SE and SE+VT/SE schemes have the largest and the second largest errors, respectively. Furthermore, as the standard deviation increases in the same range, the difference in the error between the SE/SE and SE+VT/SE schemes becomes larger.

It should be noted that the quantitative relationships among the schemes are dependent on the thresholds and parameters used in the experiment.

5.3 Relations between Subjective and Objective Assessment Results

Now we investigate the relations between the results of the subjective assessment and those of the objective assessment.

The results in Fig. 14 can be explained by those in Figs. 15, 16, and 21. First, the result in Fig. 14 that the MOS values of the schemes which perform the discarding or skipping control over the voice deteriorate largely as the standard deviation becomes larger is related to the average MU rate of voice. Next, the average MU rate of video has a similar tendency to the MOS when the standard deviation is large; that is, in this case, the SE+VT/SE scheme has the largest MOS value, and the SE+VT/skipping+SE scheme has larger MOS values than the SE+VT/skipping scheme. In addition, the result that the difference in the MOS between the SE/SE and SE+VT/SE schemes becomes larger as the standard deviation becomes larger than around 150 ms is explained by the mean square error of inter-stream synchronization.

From the above observations, we find that the MOS is closely related to the average MU rates of voice and video and the mean square error of inter-stream synchronization.

In order to confirm this, we have investigated the relation by multiple regression analysis. As a result, we have obtained the following equation:

$$V_{\mathrm{MOS}} = 0.013 R_{\mathrm{a}} R_{\mathrm{v}} - 0.0001 E_{\mathrm{inter}} - 0.65. \qquad (4)$$

V_{MOS}: Value of calculated MOS
R_{a}: Average MU rate of voice [MU/s]
R_{v}: Average MU rate of video [MU/s]
E_{inter}: Mean square error of inter-stream synchronization [ms^2]

The coefficient of determination of this equation was 0.91. In the calculation of the coefficient of determination, we considered the equation to be a linear equation by regarding $R_a R_v$ as a single variable for simplicity. To demonstrate how successfully the above equation expresses the relation, we plot the value of V_{MOS} calculated by the equation as a function of the standard deviation in Fig. 22. From this figure and Fig. 14, we find that agreement between analysis and experiment is good. Therefore, we can say that the average MU rate and mean square error of inter-stream synchronization are closely related to the MOS value.

†In [35], Steinmetz investigates the influence of constant time differences between video and voice MUs by subjective assessment in a lip synchronization experiment; that is, he assumes that there is no delay jitter. In this paper, we assess the quality of media synchronization subjectively and objectively in a systematic way under the condition that delay jitter exists.

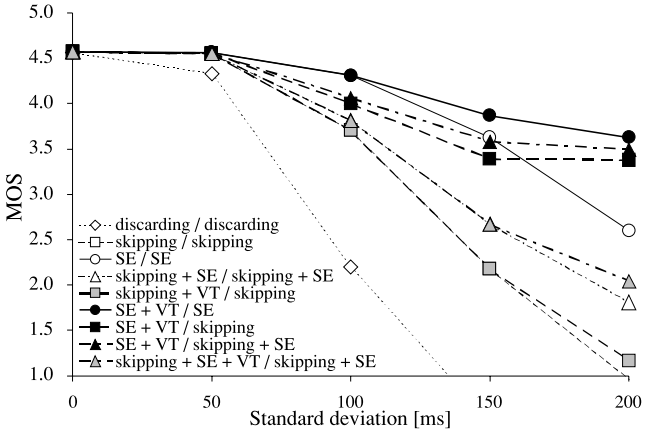

Fig. 22 MOS calculated from objective assessment results for live media.

In addition, we reached the same conclusion for stored media (see [31]).

On the other hand, it is reported that the coefficient of variation of output interval is largely dependent on the MOS value [13], [14]†. This objective measure as well as the average MU rate denotes the smoothness of output, that is, the intra-stream synchronization quality. We also examined the relation between the MOS and the coefficient of variation of output interval in this study. However, we found that the average MU rate is more closely related to the MOS than the coefficient of variation. Investigation of the reason for this difference is for further study.

6. Conclusions

This paper subjectively and objectively assessed the lip synchronization quality of nine reactive control schemes which use skipping, discarding, shortening and extension of output duration, and/or virtual-time contraction and expansion control techniques. As the assessment method, we employed computer simulation in order to get rid of the influence of the difference in implementation among the nine schemes; then, by outputting the voice and video streams actually with the output timing of MUs obtained in the simulation, we carried out subjective assessment of the quality. We also investigated the relations between the subjective assessment results and the objective ones.

As a result, we found that a scheme which uses the shortening and extension of output duration and the virtual-time contraction and expansion together for voice and the shortening and extension of output duration for video produces the best quality of lip synchronization. We also confirmed that the skipping and discarding control is not suited to voice. Furthermore, we noticed that the average MU rates of voice and video and the mean square error of inter-stream synchronization are closely related to the MOS.

Thus, we can say that the assessment method is effective. However, what kinds of measures are suited to subjective and objective assessment of media synchronization quality is not sufficiently clear yet; this is for further study.

As the next step of our research, we need to assess the interactivity as well as media synchronization quality as in [14] and [36]. We should also employ a variety of models (especially, more realistic models) for simulating network delays (for example, we may use real traces from the Internet or traces obtained by a network simulator); this is for further study. In addition, we need to investigate how the image size and the contents of voice and video sources influence the quality. Furthermore, we have to assess the quality of media synchronization of other control techniques.

†In [14], the difference in the average MU rate among schemes is negligible.

References

[1] G. Blakowski and R. Steinmetz, "A media synchronization survey: Reference model, specification, and case studies," IEEE J. Sel. Areas Commun., vol.14, no.1, pp.5–35, Jan. 1996.
[2] L. Ehley, B. Furht, and M. Ilyas, "Evaluation of multimedia synchronization techniques," Proc. Multimedia Systems'94, pp.514–519, May 1994.
[3] T.D.C. Little and A. Ghafoor, "Multimedia synchronization protocols for broadband integrated services," IEEE J. Sel. Areas Commun., vol.9, no.9, pp.1368–1382, Dec. 1991.
[4] D.P. Anderson and G. Homsy, "A continuous media I/O server and its synchronization mechanism," IEEE Computer, vol.24, no.10, pp.51–57, Oct. 1991.
[5] L. Li, A. Karmouch, and N.D. Georganas, "Synchronization in real time multimedia data delivery," Conf. Rec. IEEE ICC'92, pp.587–591, June 1992.
[6] S. Ramanathan and P.V. Rangan, "Adaptive feedback techniques for synchronized multimedia retrieval over integrated networks," IEEE/ACM Trans. Netw., vol.1, no.2, pp.246–260, April 1993.
[7] K. Rothermel and T. Helbig, "An adaptive protocol for synchronizing media streams," ACM/Springer Multimedia Systems, vol.5, no.5, pp.324–336, Sept. 1997.
[8] Y. Ishibashi and S. Tasaka, "A synchronization mechanism for continuous media in multimedia communications," Proc. IEEE INFOCOM'95, pp.1010–1019, April 1995.
[9] Y. Ishibashi, S. Tasaka, and A. Tsuji, "Measured performance of a live media synchronization mechanism in an ATM network," Conf. Rec. IEEE ICC'96, pp.1348–1354, June 1996.
[10] Y. Xie, C. Liu, M.J. Lee, and T.N. Saadawi, "Adaptive multimedia synchronization in a teleconference system," Conf. Rec. IEEE ICC'96, pp.1355–1359, June 1996.
[11] H.-Y. Chen and J.-L. Wu, "MultiSynch: A synchronization model for multimedia systems," IEEE J. Sel. Areas Commun., vol.14, no.1, pp.238–248, Jan. 1996.
[12] Y. Ishibashi and S. Tasaka, "A comparative survey of synchronization algorithms for continuous media in network environments," Proc. IEEE LCN 2000, pp.337–348, Nov. 2000.
[13] F. Kaladji, Y. Ishibashi, and S. Tasaka, "Subjective assessment of stored media synchronization quality in the VTR algorithm," IEICE Trans. Commun., vol.E82-B, no.1, pp.24–33, Jan. 1999.
[14] Y. Ishibashi, S. Tasaka, and H. Ogawa, "Subjective assessment of media synchronization quality and interactive property for live media with the VTR algorithm," Proc. ISCOM'99, pp.518–523, Nov. 1999.
[15] K. Ravindran and V. Bansal, "Delay compensation protocols for synchronization of multimedia data streams," IEEE Trans. Knowl. Data Eng., vol.5, no.4, pp.574–589, Aug. 1993.
[16] A. La Corte, A. Lombardo, S. Palazzo, and G. Schembra, "A feedback approach for jitter and skew enforcement in multimedia re-

trieval services," Conf. Rec. IEEE GLOBECOM'95, pp.790–794, Nov. 1995.
[17] M. Correia and P. Pinto, "Low-level multimedia synchronization algorithm on broadband networks," Proc. ACM Multimedia'95, pp.423–434, Nov. 1995.
[18] M.C. Yuang, B.C. Lo, Y.G. Chen, and P.L. Tien, "A synchronization paradigm with QoS guarantees for multimedia communications," Conf. Rec. IEEE GLOBECOM'99, pp.214–220, Nov. 1999.
[19] L. Qiao and K. Nahrstedt, "Lip synchronization within an adaptive QoS system," Proc. SPIE International Conference on Multimedia Computing and Networking, pp.170–181, Feb. 1997.
[20] R. Ramjee, J. Kurose, D. Towsley, and H. Schulzrinne, "Adaptive playout mechanisms for packetized audio applications in wide-area networks," Proc. IEEE INFOCOM'94, pp.680–688, June 1994.
[21] Y. Ishibashi, S. Tasaka, and T. Okuoka, "A media synchronization mechanism for MPEG video and its measured performance," Proc. 13th International Conference on Computer Communication, pp.163–170, Nov. 1997.
[22] S. Tasaka, H. Nakanishi, and Y. Ishibashi, "Dynamic resolution control and media synchronization of MPEG in wireless LANs," Conf. Rec. IEEE GLOBECOM'97, pp.138–144, Nov. 1997.
[23] I. Kouvelas, V. Hardman, and A. Watson, "Lip synchronization for use over the Internet: Analysis and implementation," Conf. Rec. IEEE GLOBECOM'96, pp.893–898, Nov. 1996.
[24] S.B. Moon, J. Kurose, and D. Towsley, "Packet audio playout delay adjustment: Performance bounds and algorithms," ACM/Springer Multimedia Systems, vol.6, no.1, pp.17–28, Jan. 1998.
[25] M. Kato, N. Usui, and S. Tasaka, "Performance evaluation of stored media synchronization in PHS," IEICE Trans. Commun. (Japanese Edition), vol.J80-B-II, no.9, pp.749–759, Sept. 1997.
[26] M. Kato, N. Usui, and S. Tasaka, "Media synchronization control based on buffer occupancy for stored media transmission in PHS," IEICE Trans. Fundamentals, vol.E81-A, no.7, pp.1378–1386, July 1998.
[27] L. Lamont, L. Li, R. Brimont, and N.D. Georganas, "Synchronization of multimedia data for a multimedia news-on-demand application," IEEE J. Sel. Areas Commun., vol.14, no.1, pp.264–278, Jan. 1996.
[28] S. Jha and M. Fry, "Continuous media playback and jitter control," Proc. IEEE Multimedia Systems'96, pp.245–252, June 1996.
[29] S.T. Liang, P.L. Tien, and M.C. Yuang, "Threshold-based intra-video synchronization for multimedia communications," IEICE Trans. Commun., vol.E81-B, no.4, pp.706–714, April 1998.
[30] M.C. Yuang, P.L. Tien, and S.T. Liang, "Intelligent video smoother for multimedia communications," IEEE J. Sel. Areas Commun., vol.15, no.2, pp.136–146, Feb. 1997.
[31] Y. Ishibashi, S. Tasaka, and H. Ogawa, "A comparison of media synchronization quality among reactive control schemes," Proc. IEEE INFOCOM 2001, pp.77–84, April 2001.
[32] S. Tasaka, T. Nunome, and Y. Ishibashi, "Live media synchronization quality of a retransmission-based error recovery scheme," Conf. Rec. IEEE ICC 2000, pp.1535–1541, June 2000.
[33] D.L. Mills, "Internet time synchronization: The network protocol," IEEE Trans. Commun., vol.39, no.10, pp.1482–1493, Oct. 1991.
[34] ITU–R 500–5, "Method for the subjective assessment of the quality of television pictures," International Telecommunication Union, Sept. 1992.
[35] R. Steinmetz, "Human perception of jitter and media synchronization," IEEE J. Sel. Areas Commun., vol.14, no.1, pp.61–72, Jan. 1996.
[36] T. Kurita, S. Iai, and N. Kitawaki, "Effects of transmission delay in audiovisual communication," IEICE Trans. Commun. (Japanese Edition), vol.J78-B-I, no.4, pp.331–339, April 1993.
[37] Y. Ishibashi, S. Tasaka, and T. Inoue, "Joint synchronization between live and stored media in network environments," Proc. IEEE ICCCN 2000, pp.442–446, Oct. 2000.

Yutaka Ishibashi received the B.S., M.S., and Ph.D. degrees from Nagoya Institute of Technology, Nagoya, Japan, in 1981, 1983, and 1990, respectively. From 1983 to 1993, he was with NTT Laboratories. In 1993, as an Associate Professor, he joined Nagoya Institute of Technology, in which he is now a Professor in the Department of Computer Science and Engineering, Graduate School of Engineering. From June 2000 to March 2001, he was a Visiting Professor in the Department of Computer Science and Engineering at the University of South Florida. His research interests include networked multimedia applications, media synchronization algorithms, and multimedia communication protocols. Dr. Ishibashi is a member of the IEEE, ACM, Information Processing Society of Japan, and the Institute of Image Information and Television Engineers.

Shuji Tasaka received the B.S. degree in electrical engineering from Nagoya Institute of Technology, Nagoya, Japan, in 1971, and the M.S. and Ph.D. degrees in electronic engineering from the University of Tokyo, Tokyo, Japan, in 1973 and 1976, respectively. Since April 1976, he has been with Nagoya Institute of Technology, where he is now a Professor in the Department of Computer Science and Engineering, Graduate School of Engineering. In the 1984–1985 academic year, he was a Visiting Scholar in the Department of Electrical Engineering at the University of California, Los Angeles. His current research interests include wireless networks, multimedia QoS, and multimedia communication protocols. He is the author of a book entitled *Performance Analysis of Multiple Access Protocols* (MIT Press, Cambridge, MA, 1986). Dr. Tasaka is a member of the IEEE, ACM, and Information Processing Society of Japan.

Hiroki Ogawa received the B.S. and M.S. degrees from Nagoya Institute of Technology, Nagoya, Japan, in 1999 and 2001, respectively. He joined NTT DATA Corp. in April, 2001. He was engaged in research on quality assessment of media synchronization at Nagoya Institute of Technology.