PAPER Special Section on New Challenge for Internet Technology and its Architecture

# Performance Monitoring of VoIP Flows for Large Network Operations

**Yoshinori KITATSUJI**[†,††a)], **Satoshi KATSUNO**[†], **Katsuyuki YAMAZAKI**[†††], **Masato TSURU**[††††], ***Members***, ***and*** **Yuji OIE**[††††], ***Fellow***

**SUMMARY** The monitoring of performance in VoIP traffic has become vital because users generally expect VoIP service quality that is as high as that of PSTN services. A lightweight method of processing by extracting VoIP flows from Internet traffics is proposed in this paper. Estimating delay variations and the packet loss ratio using knowledge about specific features and the characteristics of VoIP flows, i.e., the inter-packet gap (IPG) which is constant in VoIP flows, is also proposed. Simulation with actual traffic trace is used to evaluate the method, and this revealed that delay variations (IPG variance) can be accurately estimated by monitoring only a few percentage of all flows. The proposed method can be used as a first-alert tool to monitor large amounts of flows to detect signs of degradation in VoIP flows. The method can be used by ISPs to estimate whether VoIP flow performance is adequate within their networks and at ingress from other ISPs.

***key words:*** *performance monitoring, VoIP, IP flow, delay variation, and inter-packet gap*

## 1. Introduction

Internet service providers (ISPs) are now facing the challenge of providing reliable transport for real-time traffic, e.g., voice over IP (VoIP) and video conferencing, due to recent advances in broadband Internet access. Many ISPs now provide VoIP services which include interconnection to the public switch telephone networks (PSTNs), as well as Internet accessibility. In Japan, 14% of home customers signing contracts with ISPs check if VoIP services are included and 43% of home customers use VoIP services [1]. Most users generally expect VoIP-service quality to be as high as that of PSTN services. Therefore, ISPs must now seriously consider VoIP-traffic performance when designing and operating their networks.

While the Internet has long carried various types of traffic (e-mail, the Web, and peer-to-peer), VoIP traffic is quite different from that previous. It is difficult to identify degradation in VoIP from the utility information of a link since the link has different types of traffic. For example, VoIP performance may be degraded due to the burstiness of other traffic even if link utility is low. Therefore, the ability to monitor VoIP traffic exclusively is of great importance.

First-alert tools are generally used for ordinary network monitoring. They passively monitor traffic to detect signs of trouble, i.e., congestion, network attacks, or intrusions and they warn network operators if the signs have been detected. Once network operators detect an unusual situation a variety of actions is taken with exclusive tools. Although the ISPs have so far been relying on link-utility information, e.g., traffic rate variation monitored by MRTG [2], an exclusive first-alert tool is now required to monitor the performance of VoIP traffic.

An ISP must monitor the performance of VoIP traffic at ingress and egress to know whether its performance is adequate within the network, and whether this is adequate at the ingress connected to customers. Monitoring performance at interconnections to other ISPs is also important when the VoIP services are provided with the ISPs. This is because an ISP should know whether VoIP performance has been degraded within its own network or is already degraded when VoIP packets enter its network.

The key metrics of VoIP performance are delay variation and packet loss. The inter-packet gap (IPG), which is the time delay between successive packets observed in individual VoIP flows, is used to estimate delay variation. We define a flow as a sequence of packets distinguished by five tuples that is, IP addresses, protocols and port numbers with an IPG less than the given timeout, in this paper. Detecting the missing packet in losses, i.e., counting the lack of numbered packets in individual flows by tracking the application data in a packet, complicates packet processing in counting packet loss. Because of this, a VoIP-flow performance monitoring system often can not deal with large amounts of flows. To solve this problem, we propose a method of estimating delay variation and packet losses in VoIP flows by using features, such as, fixed packet length and almost fixed IPG, to detect VoIP flow to solve this problem.

NetFlow [3] and sFlow [4] are monitoring tools that produce high-level statistics on flows identified with five tuples. Since they are generally implemented in routers and switches, computing the statistics can not obtain sufficient processing power for them. Therefore, their monitoring involves dealing with the limited number of flows or provides the severely inaccurate statistics due to limited packet sampling.

Manuscript received January 26, 2007.
Manuscript revised April 23, 2007.

[†] The authors are with KDDI R&D Laboratories, Inc., Fujimino-shi, 356-8502 Japan.
[††] The author is with Kyushu Research Center, NICT, Kitakyushu-shi, 802-0001 Japan.
[†††] The author is with Nagaoka Universisty of Technology, Nagaoka-shi, 940-2188 Japan.
[††††] The authors are with the Department of Computer Science and Electronics, Kyushu Institute of Technology, Iizuka-shi 820-8502 Japan.
a) E-mail: kitaji@kddilabs.jp

DOI: 10.1093/ietcom/e90–b.10.2746

Calyptech produces an all-in-one system to non-intrusively monitor the quality of voice exchanged in various types of network, such as, VoIP, Wireless, PSTN, and enterprise networks [5]. The VoIP network's product, M10, collects VoIP packets with a speed of up to 1 Gbit/s and analyzes the mean opinion score using P.SEAM (ITU-T P.563 [6]), E-Model (ITU-T G.107 [7]), echo return loss, periods of non-speech, and noise. Most metrics are for voice quality degraded in terminal adapters (TAs) or phones in a VoIP system and it does not directly monitor metrics representing the network properties affecting VoIP flows. These metrics are of great importance to the network operations. The complexity of analysis probably also limits the number of flows that must be monitored in the backbone network.

Deri proposed the design and implementation of an open-source tool for detecting and measuring VoIP traffic [8]. It collects NetFlow-like information on VoIP flows with nProbe [9] and analyzes their metrics with ntop [10] to enable the network properties affecting VoIP flows to be understood. The tool is targeted at monitoring various types of VoIP flows, such as, those with variable and fixed bit rates and packet sizes. Such flexibility results in obstacles to enabling the tool to deal with large numbers of flows in high-speed links, e.g. those at 10 Gbit/s.

We emphasize that the scalability of large numbers of VoIP flows is more important than flexibility for their various types in ISP-network operations. Therefore, we propose a lightweight method of distinguishing VoIP packets to attain this scalability.

The remainder of this paper is organized as follows. Section 2 addresses VoIP-flow monitoring problems in terms of scalability. Section 3 discusses our analysis of traffic traces obtained from Abilene [11] and a certain ISP in Japan. Section 4 describes our method of extracting VoIP flows, and estimating delay variations and packet loss on the basis of the feature identified from VoIP flows. Section 5 discusses the evaluation of our method through packet-based simulation with traffic traces. Section 6 we discusses our application of the new method to a real network to monitor operations. Section 7 summarizes the paper and contains some concluding remarks.

## 2. Monitoring Performance of VoIP Flows

### 2.1 Monitoring Locations in Backbone Network

Figure 1 outlines the location of tools used to monitor VoIP flows in an ISP's backbone network. Monitor *A* monitors VoIP flows from customers and checks if degradation has occurred in a segment between the customers and the monitoring point. Degradation occurring in the backbone network is monitored in a segment between any pair of *A*, *B* and *C*, or *D*.

Degradation in the core is more accurately inferred than that in customers since characteristics of VoIP flows, such as, IPG and the number of lost packets, monitored between two monitoring points can be compared. Because monitoring detects the degradation occurred in the customer on the basis of a single point monitoring, the features of packet generation of a terminal adapter, such as, almost the same inter-packet gap over packets in VoIP flows (described in Sect. 2.2), should be used. Such uses are the focus of this paper. Section 3 demonstrates that the actual features of VoIP flows are usable.

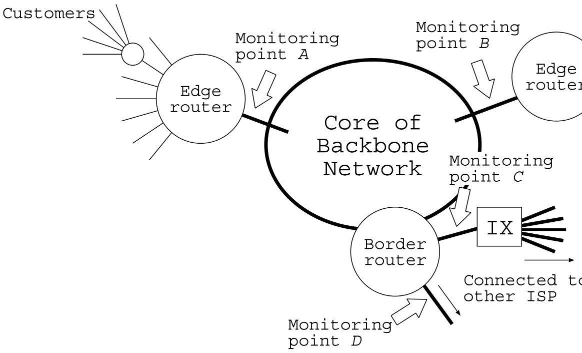

**Fig. 1** Location of monitoring tools in backbone network. (IX is Internet exchange point)

### 2.2 Identification of VoIP Flows

A VoIP system is generally composed of an SIP server, VoIP TAs, and phones. A call made over the Internet is set up using the following steps.

1. A caller TA requests a SIP server to setup a call with information on the caller TA when a number is dialed.
2. The SIP server announces the information to set up the connection on the caller TA, i.e., an IP address and the dynamically assigned port number to the callee TA, collects the information on the callee TA, and responds to the caller TA with the callee information.
3. The caller TA establishes a connection with the callee TA and a bidirectional VoIP flow begins.

As we can see from this procedure, the monitoring tool needs to infer the port numbers of a VoIP flow, or to collect these from the VoIP entities to monitor them.

The SIP server and TAs can potentially notify of IP addresses and dynamically assigned port numbers for each of the VoIP flows. However, the SIP server or TAs and the monitoring tool do not share a common communication protocol at present. Therefore, we assumed that the monitor would identify VoIP flows without the information given by the VoIP entities.

There are four basic steps in monitoring the performance of VoIP flows:

1. capturing packets,
2. distinguishing VoIP packets,
3. identifying a flow from the VoIP packet using five tuples, and
4. estimating the metrics of the VoIP flow.

Three types of method are generally used for step 2: 1) port numbers in the UDP header, 2) a flag marked by the TA, (i.e., a bit flag marked in an IP header), and 3) attribute data used by the VoIP application. However, 1) can not be used because the VoIP entities are assumed no to have been informed of the dynamic assigned port numbers for VoIP. Although the VoIP packets for 2) are distinguished without error, only a few ISPs that can modify the TA they distribute to customers can use flags. Finally, identification using 3) is not always accurate because the attributes can be altered or concealed by VoIP applications [12].

We propose a method that distinguishes VoIP packets using the features of packets in VoIP flows in this paper. Our method can greatly reduce the workload in identifying flows. Therefore, it enables traffic as fast as 10-Gbit/s to be monitored. Although the method can only be applied to limited flows from VoIP applications using a specific codec, we will demonstrate such a VoIP application can easily be adopted for VoIP services in ISPs.

### 2.3 Metric Representing Degradation of VoIP Flows

The key metrics of VoIP performance are delay variations and the rate of consecutive packet loss. Delay variations can reveal the extent of congestion occurring in network equipment. The rate of packet loss can reveal the frequency of heavy congestion and failure occurring in links and network equipments including TAs.

The monitor may be able to accurately determine delay variations influencing a TA if it can infer and emulate the process of packet arrival in a TA receiving a VoIP flow. However, such monitoring is inadequate in ISPs with large-scale networks because there should be large amounts of flows. Although information on performance conveyed by the extended report in the RTP header [13] in a VoIP packet may be used, Sect. 3 explains that most VoIP packets do not carry performance reports (as determined from the packet size). For these reasons, we used the variance in IPG showed by VoIP flows as delay variations. Since IPG also increases as the number of consecutive lost packets increases, we defined the corrected IPG as the delay variations, as described in Sect. 4.

The packet sequence in a VoIP flow must be tracked in single-point monitoring to determine the rate at which consecutive packets are lost. Although the IP header has an identifier field in which sequence number is often carried, a packet number cannot be sequentially assigned to each packet of a flow when another flow occurs simultaneously. Although the RTP header can have a sequence number, it is also conveyed in the optional header and can not be checked in most VoIP packets as discussed in the analysis in Sect. 3.

Therefore, we estimate the rate of consecutive lost packets from the features of an almost fixed IPG, as will be explained in Sect. 4. This enables single-point monitoring to determine the rate of packet loss without using the sequence number carried in the packet header.

## 3. Characteristics of VoIP Flows

This section discusses the characteristics of VoIP flows derived from intuition and then verifies these with traffic traces. We then identify the best parameters for extracting VoIP flows.

Intuitively, VoIP flows last for prolonged periods and have fixed packet length and fixed IPG.

- A Japanese Government white paper [1] reported that an average business call lasts for 39 s and an average home call lasts for 92 s. The duration of a VoIP flow should be on the order of tens of seconds. Although the white paper gave no statistics on VoIP flows, it is difficult to conceive that VoIP calls are significantly shorter than ISDN and mobile phone calls.
- There should be a large number of VoIP flows with a fixed packet length, because it is likely that a G.711 (64 Kbit/s) codec [14] will be adopted for most VoIP applications to enable a TA to interconnect to PSTNs, and because the increased availability of broadband Internet access has eliminated concerns about bandwidth use.
- Most TAs are designed to periodically generate packets, even when a period of silence during calls to simplify the time-synchronization mechanism used to decode the digital signal from received packets.

We analyzed actual traffic to verify these suppositions.

Figure 2 plots cumulative distributions for the average packet length of individual UDP flows for the traffic traces collected from a 2.4-Gbit/s link in Abilene in August 2002 and from the 1-Gbit/s link of an access point to the backbone of an ISP in Japan in July 2003. The traffic was fully captured for 100 s in Abilene, and for 60–70 s for the ISP in Japan (Table 1). The distributions are for flows lasting more than certain durations (5, 10, 15, and 20 s).

The distributions for calls that lasted more than 10–20 s were similar for the ISP in Japan. More than 45% of the

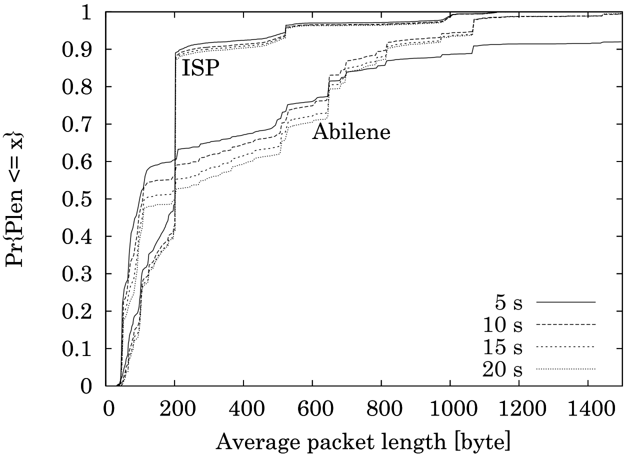

**Fig. 2** Cumulative distribution of average packet length for UDP flows of more than certain durations.

flows had an average packet length of 200 bytes. A packet encoded with the G.711 (64-Kbit/s) codec carried 160 bytes of voice data (64000 × 0.02/8) assuming it took 0.02 s to become packetized. Hence, an IP packet with a 20-byte IP header, an 8-byte UDP header, and a 12-byte RTP header is 200 bytes long. Therefore, flows with 200-byte packets are expected from TAs with the G.711 codec. The distributions for Abilene are not similar to the ISP's. More than 20% of the flows had an average packet length of 50 bytes. The IP packets may carry 10-byte voice data that arrive from the G. 729 (8 Kbit/s) codec [15], if packetization takes 0.01 s.

We analyzed cumulative distributions of the average IPG for UDP flows of packets that lasted more than 10 s, which were about 200-byte long for ISP or 50 for Abilene (Fig. 3), to determine the relationship between codec and IPG (a packetizing interval). The plots show that the 50-byte packets did not arrive from the G.729 codec because average IPGs had mostly unexpected values (more than half of the IPGs were about one second). We focused on the ISP's traffic traces after this because most of the flows extracted from Abilene did not have the features of a fixed IPG. The ISP's traffic trace also showed that more than 90% of flows had an average IPG of about 0.02 s. The slight variations from 0.02 s were probably caused by flows being attracted by delay variation.

We last analyzed the duration of flows longer than 10 s with 200-byte-long packets and an IPG ranging from 0.01 to 0.03 s (no figure). The average duration was 47 s and more than 75% of the flows lasted longer than 30 s and 35% lasted longer than 60 s. Because they reduce frequency at which VoIP flows must be registered in the monitoring tool (see pseudo code in Sect. 4.2), long flows are easier for the proposed method to handle.

**Table 1** Characteristics of traffic traces.

| Source | Length of data set (s) | Link bandwidth | No. of data sets | Total no. of packets |
|---|---|---|---|---|
| Abilene | 100 | 2.4 Gbit/s | 12 | 1.26 bill. |
| ISP | 60 – 70 | 1 Gbit/s | 55 | 175 mill. |

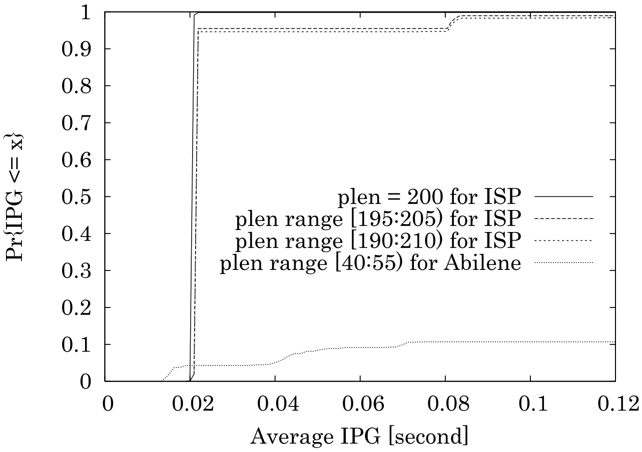

**Fig. 3** Cumulative distribution of average IPG for UDP flows lasting more than 10 s in range of packet lengths.

Therefore, we used flows with 200-byte-long packets and an average IPG of around 0.02 s with our method. While many VoIP applications currently generate packets of different lengths and IPGs, it is likely that they will soon generate fixed-length packets at a fixed rate for the reasons described earlier in this section. When this happens, the new method we propose can easily be applied merely by adjusting the packet-length and IPG conditions.

## 4. Proposed Method

As described in Sect. 2.3, the consecutive lost packets and the variance in IPG were estimated to represent the performance of VoIP flows. The IPGs were collected from all packets in each of some amounts of flows monitored during a certain period (e.g., 0.1 s) at a monitoring point. There is a trade-off between the accuracy of degradation that rises as the number of monitored flows is increased and the workload to compute this degradation. The actual number (or ratio) of monitored flows in all flows is an operational issue. Although accurate estimates are required to find even small degradation in VoIP flows, this may result frequent alarms as the entire traffic as well as VoIP flows change. However, inaccurate monitoring may neglect serious degradation in VoIP flows.

This paper discusses the relationship between the amount of monitored flows and accuracy with the evaluation results presented in Sect. 5. We believe that our proposed metrics is not affected by how flows are selected. Therefore, we present a simple yet concrete method of selecting flows (greedily using $N$ first arrival flows) in Sect. 4.2.

### 4.1 Estimates of Delay Variations and Consecutive Lost Packets

When a longer than an ideal IPG (based on the features of VoIP flows as described in Sect. 3) is found, we assume that there must be a single or consecutive multiple packet loss. We then correct the IPG that rises when there are lost packets, as if there were an adequate number of packets in the long IPG.

Let $t^i_j$ denote the time when packet $j$ in flow $i$ passes the monitoring point, and $x^i_j$ denote the IPG between the $j$-th and $(j+1)$-th packets. We define a pseudo-IPG variation, denoted by $y^i_j$ $(> 0)$, to represent the difference between an ideal IPG, denoted by $\tilde{X}$, and an actual IPG $x^i_j$. We let $y^i_j$ exclude large IPG occurring when previously arriving consecutive packets are lost. We use a factor of the ideal IPG, denoted by $k^i_j$ $(= \max(1, \lfloor x^i_j/\tilde{X} + \alpha \rfloor))$ to define $y^i_j$, which satisfies

$$k^i_j = \begin{cases} 1 & \text{if } 0 \le x^i_j < (1+\alpha)\tilde{X} \\ n & (n \ge 2) \\ & \text{if } (n+\alpha-1)\tilde{X} \le x^i_j < (n+\alpha)\tilde{X}. \end{cases} \qquad (1)$$

$y^i_j$ is defined with $k^i_j$ as,

$$y^i_j = |x^i_j - k^i_j\tilde{X}|\,. \qquad (2)$$

Parameter $\alpha$ (termed the *receiver-TA buffer factor*) determines how much buffering delay is expected in the receiver TAs. This parameter can be used to estimate how large the IPG of a packet can be considered as packet loss in terms of the receiver TA, e.g., even a small IPG should be considered as a packet loss when a receiver TA's buffering delay ($\alpha$) is short. The recommended value for this parameter is 0.5 through 2.5 which is congruent with the 0.01 through 0.05 s in jitter in the case of G. 711 codec (IPG of 0.02 s). Although the buffering delay in a receiver-TA should absorb all the delay variations occurring between the sender and receiver TAs in the ETSI recommendations [16], long buffering delay is not recommended because it increases the delay to decoding packets. The receiver-TA's buffer delay should also be defined with the network operations taking their service-level policy into account.

We can call $k^i_j$ an inferred loss because, with an indicator function

$$\chi_{k^i_j \geq 2} = \begin{cases} 0 & \text{if } 0 \leq k^i_j < 2 \\ 1 & \text{if } k^i_j \geq 2, \end{cases} \qquad (3)$$

The rate of long IPGs treated as the consecutive packet loss ($x^i_j \geq (1+\alpha)\tilde{X}$) is represented as

$$r = \frac{\sum_{i,j}\chi_{k^i_j \geq 2}}{N\sum_i n_i}, \qquad (4)$$

where $N$ is the number of monitored flows and $n_i$ is the number of packets in flow $i$ within a given monitoring interval (e.g., 0.1 s).

$r$ is considered as the rate of consecutive packet loss when the buffer size of the links through which the monitored packets have passed is $\alpha\tilde{X}$. In addition, when the buffer size is larger than $\alpha\tilde{X}$, $r$ is also considered as the rate for the combination of consecutive packet loss and IPGs exceeding $(1+\alpha)\tilde{X}$. The proposed method in the latter may count packet loss for long-IPG packets. However, if an appropriate value is assigned to $\alpha$, such a long IPG will probably degrades voice quality since it is considered to be an effective loss. Therefore, we define $r$ as an estimator of the rate of consecutive packet loss in terms of VoIP flow quality. After this, we term $r$ *degraded packet rate*. With this estimator, the packet loss rate can be estimated even with single-point monitoring.

We define delay variation, denoted by $S^2$, as the variance in pseudo-IPGs,

$$S^2 = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{n_i}\sum_{j=1}^{n}{y^i_j}^2\,. \qquad (5)$$

This enables estimates of the delay variations from which the impact of packet loss has been removed when there are frequent packet losses. After this, we term the standard deviation of corrected IPGs ($S$) the *IPG deviation*. This estimator enables network operators to segregate two types of degraded VoIP flows, i.e., when 1) congested links increase delay variations and cause packet loss if they are heavily blocked up, and when 2) processing failures by routers and switches, or large bit error in links cause packet loss but do not increase the delay variation. The degraded packet rate and IPG deviation increase in the first case. However, in the second case, only the degraded-packet rate increases.

### 4.2 Algorithm to Register Monitored VoIP Flows

We found from the observation described in Sect. 3 that there was a sufficiently large number of flows with packets of 200 bytes and an average IPG of around 0.02 s. We therefore extracted flows having 200-byte packet lengths, 0.01–0.03-s average IPGs, and durations of more than 10 s to monitor performance. The algorithm verifies the mean of IPGs for a monitored flow, after which the flow lasts more than 10 s.

Figure 4 shows a pseudo-code used to identify monitored VoIP flows. The algorithm greedily registers VoIP flows in set $\mathcal{M}$, until the number of registered flows reaches a given number $N$. Identified flows lasting less than 10 s are registered in set $\mathcal{P}$. The algorithm first checks if the packet is UDP and 200-byte long when a packet arrives, and identifies a flow of packets satisfying these conditions. The process to

```
Function $MonitoringFlow$;
var $N$: Number of flows to be monitored;
var $\tilde{X}$: Expected average IPG;
var $L_{min}$: Minimum packet length;
var $L_{max}$: Maximum packet length;
var $T$: Threshold to register flows;

begin
  var $\mathcal{M} := \emptyset$: set of monitored flow;
  var $\mathcal{P} := \emptyset$: set of active flow w/t short duration;
  var $\mathcal{F} := \emptyset$: set of flow information base (FIB);
  repeat
    var $p := ArrivingPacket(L_{min},\ L_{max})$;
    var $f := FlowID(p)$;
    $UpdateFIB(p,\ f,\ \mathcal{F})$;
    if $f \in \mathcal{M}$ then
      $UpdateStatistics(f,\ \mathcal{M},\ \mathcal{F})$;
    else
      if $f \in \mathcal{P}$ then
        if $FlowDuration(f,\ \mathcal{F}) \geq T$ then
          $RemoveFlow(f,\ \mathcal{P})$;
          if $\tilde{X}/2 \leq IPG(f) < 3\tilde{X}/2$ then
            $InsertFlow(f,\ \mathcal{M})$;
            $UpdateStatistics(f,\ \mathcal{M},\ \mathcal{F})$;
          endif;
        else
          $AccountPacketAndByte(p,\ f,\ \mathcal{F})$;
        endif;
      else
        if $|\mathcal{M}| + |\mathcal{P}| < N$ then
          $InsertFlow(f,\ \mathcal{P})$;
        endif;
      endif;
    endif;
    $RemoveInactiveFlow(\mathcal{M},\ \mathcal{P},\ \mathcal{F})$;
  until receiving a signal to terminate;
end;
```

**Fig. 4** Flow-monitoring algorithm. Parameters, $N$, $\tilde{X}$, [$L_{min}$, $L_{max}$), and $T$, are set to 100, 0.02 s, 200 or 201 bytes, and 10 s, for example.

identify a flow from a packet is inevitable because to obtain the IPG which is acquired from a gap between consecutive packets in a flow. If the packet is UDP and 200-byte long, the IPG variance is computed as described in Sect. 4.1, when the flow is already registered in set $\mathcal{M}$. The flow is moved to set $\mathcal{M}$ from set $\mathcal{P}$, if the flow is found in $\mathcal{P}$, lasts more than 10 s and has an average IPG, i.e., $IPG(f)$ within $\tilde{X}/2$ and $3\tilde{X}/2$. The algorithm also periodically monitors the termination of flows registered in sets $\mathcal{M}$ and $\mathcal{P}$, and removes them if they terminate.

## 5. Evaluation of Performance

The accuracies of degraded-packet rate and IPG deviation were first verified through the simulating the packet-arrival process to evaluate the estimated the rate of consecutive packet loss and the delay variations. We then found why their estimation accuracies of estimation were different. We also found that our proposal could effectively be used for a first alert tool even if packet-loss assignment contained error.

### 5.1 Accuracy

The experiment simulated the ISP traffic traces used in Sect. 3, and virtual VoIP flows with 200-byte packet lengths, and an ideal IPG (0.02-s) forwarded through a single virtual link. A traffic trace that changed its bit rate most rapidly was used for this simulation. We will call the virtual VoIP flow the *inserted flow* after this. The virtual link had a 0.01-s buffer size and a specific bandwidth of less than 1 Gbit/s (actual link bandwidth from which the trace data was collected) to cause queuing delay in packets. Therefore, a large IPG in a flow leading to increased degraded-packet rate represents actual packet loss since the buffer size is sufficiently small.

We focused on how accurately the degraded-packet rate represented the actual rate of consecutive packet loss and how accurately the IPG deviation represented the standard deviation in the actual IPG. Through simulation, we assigned 0.5 to $\alpha$. The accuracy of the proposed method may change as this value changes. The relationship between this value and the accuracy of estimation, and an appropriate value for $\alpha$ to satisfy the rigor required in monitoring, taking the degradation level perceived by the receiver TA into consideration remain as future work.

Figure 5 shows a Q-Q plot of the actual packet-loss rate for all inserted flows ($X$-axis) and the degraded-packet rate for some amounts of flows indicated by the symbols ($Y$-axis). From 100 to 500 flows were used for monitoring a thousand inserted flows (80 Mbit/s in total). The virtual link was shaped to 381 Mbit/s, resulting in a total utility rate of 95% (The traffic traces averaged 282 Mbit/s, and the ideal flows amounted to 80 Mbit/s).

The degraded-packet rate as seen from the figure is not very accurate. It is 0.5–2 times the actual packet loss rate in the 0–0.2% range. The degraded-packet rate is often overestimated including error of 0.1–0.2% in ranges where the actual packet-loss rate is larger than 0.2%. This is because the proposed method cannot detect the loss in all of the inserted flows. Additional experiments with 500, 2000, and 5000 inserted flows (no figures), demonstrated that accurate estimates of packet loss required at least 20–50% of all inserted flows to be monitored and that the estimates also included error of 0.1–0.2%.

This means that huge numbers of probes need to be inserted into the monitored segment to estimate the consecutive packet-loss rate, when actively measuring with probes that emulate VoIP flows. Since the influence of such huge amount of probe traffic on actual traffic is unavoidable, passive measurement is much better suited to monitoring the performance of VoIP flows than active.

We next estimated delay variations. Figure 6 shows a Q-Q plot for the standard deviation of actual IPGs ($X$-axis) and the IPG deviation ($Y$-axis) for all inserted flows. The conditions for selected traffic traces, the number of inserted ideal flows, buffer size, and the bandwidth of the virtual link

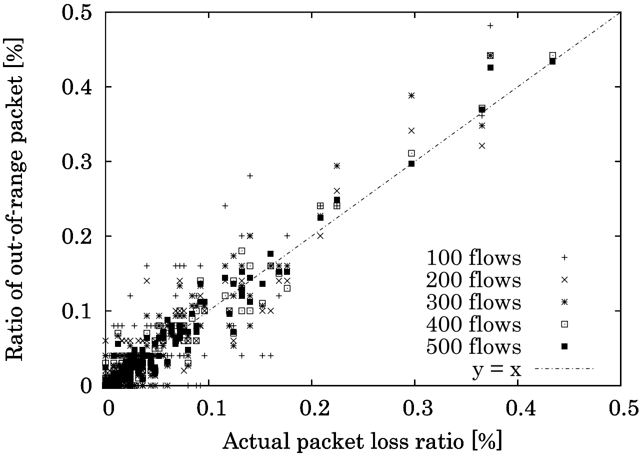

**Fig. 5** Q-Q plot of actual packet loss rate for all inserted flows ($X$-axis) and degraded packet rate of monitored flows ($Y$-axis) when buffer size was limited to that at which maximum queuing delay was 0.01 s.

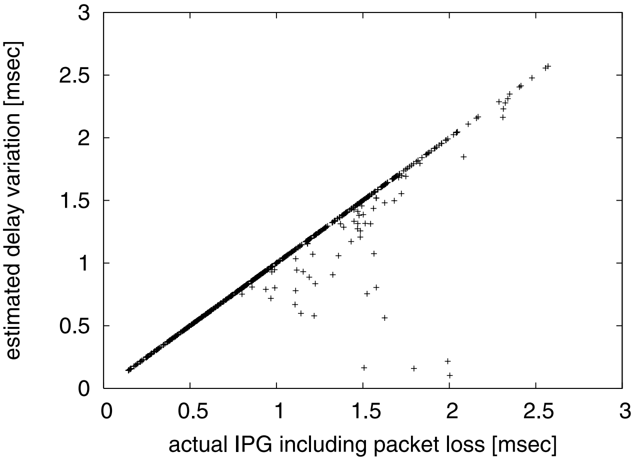

**Fig. 6** Q-Q plots of standard deviation for actual IPGs ($X$-axis) and IPG deviation ($Y$-axis) for all inserted flows.

are the same as those to evaluate the degraded loss rate. The figure shows that almost all IPG deviations are accurate but some of these are much lower than the standard deviation of the actual IPG. The reason for this is that the corrected IPG does not include a long IPG, which is observed when a packet arrives after consecutive packet are lost.

The result in Fig. 7 helped us to confirm this hypothesis. The figure shows a Q-Q plot of the standard deviation of actual IPGs (*X*-axis) excluding long IPGs, which lead to consecutive packet loss and IPG deviations (*Y*-axis) for all inserted flows. Both values are fairly similar over all delay variations (0.1 through 2.5 ms). The IPG deviations can accurately present the actual delay variations without precise information on how much packet loss has occurred in each monitored flows.

The IPG deviation for all inserted flows (*X*-axis) and that for a certain number of flows (*Y*-axis) were compared (Fig. 8) to evaluate the accuracy of IPG deviations from a certain number of VoIP flows (not all the flows). The figure shows that estimates gave fairly accurate estimation despite that only 20 though 80 flows in a thousand one are monitored. The estimate from 80 flows was slightly more accurate.

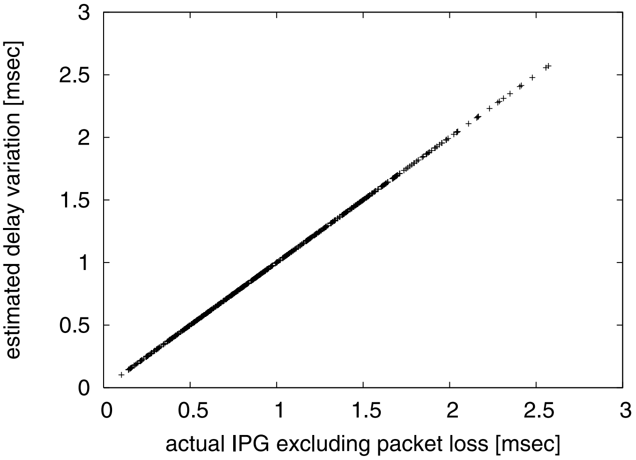

**Fig. 7** Q-Q plots of standard deviation for actual IPGs without large IPG (*Y*-axis) to which leads to consecutive packet loss (*X*-axis) and IPG deviation (*Y*-axis) for all inserted flows.

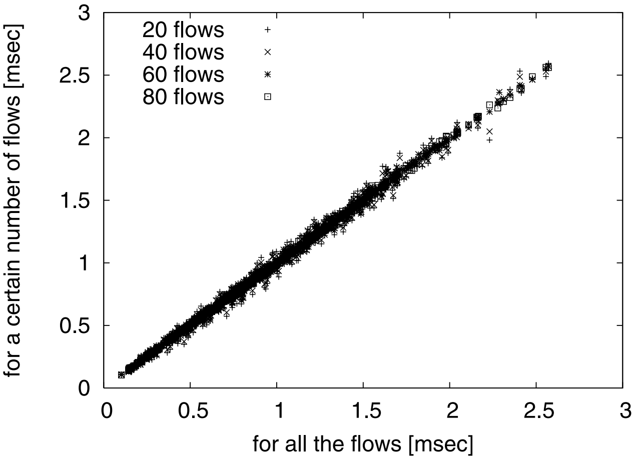

**Fig. 8** Q-Q plots of average difference from ideal IPG for all inserted flows (*X*-axis) and that for certain numbers of monitored flows (*Y*-axis).

This proves that IPG deviations are a sufficiently accurate reflection of the actual delay variations represented by the IPG, and that IPG deviations from a few percent of flows (2%) can accurately represent those from all VoIP flows.

### 5.2 Number of Monitored Flows and Accuracy

We found from the simulation, that the number of monitored flows $N$ should be sufficiently large to enable the degraded-packet rate to be accurately estimated. However, the estimated delay variations are fairly accurate, even if the percentage of monitored flows is 2% of all flows.

If the distribution of the rate of consecutive packet loss and IPG deviations over all flows in each monitoring interval follows a normal distribution, a metric with small variance in the distribution should give more accurate estimates even if small amounts of flows are monitored. Table 2 lists the relative standard deviations (standard deviation normalized by average) over all flows for the actual rate of consecutive packet loss and IPG excluding consecutive packet loss. The statistics (minimum, average, and maximum) were computed over all monitoring intervals. The table shows that large amounts of flows are required to establish the level of the consecutive packet loss since its distribution over flows has quite a large relative standard deviation. However, the IPG can be estimated with small amounts of flows due to its much smaller relative standard deviation.

Although it is still assumed that each metric over all flows follows a normal distribution, these results strongly suggest that the delay variations have a higher correlation than consecutive packet loss over flows sharing the same queue.

### 5.3 Incorrect Assessment of Packet Loss and Applicable Limitations

The proposed degraded packet-loss rate includes incorrect assessments of packet loss in the following situations.

1. When the latter of two consecutive packets has long delay, it can be judged as lost despite the fact that all the link buffers are sufficiently large and this latter packet even arrives at the receiver TA.
2. When there are three consecutive packets where the first packet has long delay, the second is lost, and the third is not delayed relative to the first, the loss of the

**Table 2** Relative standard deviation of each metric over all flows in each monitoring interval (0.1 s) for delay variation and consecutive packet losses.

| Metric | Minimum | Average | Maximum |
|---|---|---|---|
| Consecutive packet loss rate | 4.522 | 15.72 | 31.61 |
| IPG | 0.00072 | 0.00601 | 0.02507 |

second may be overlooked.

Both cases reveal the outcome with the proposed method if it is used to assess the packet loss rate. However, both are allowable in terms of use for first-alert monitoring. In the first case, packets with long delay are often effectively lost in terms of the receiver TA, as described in Sect. 4.1. Therefore, such over-calculation of packet loss is permissible to detect degradation in the performance of VoIP applications.

There are few overlooking in the case of three consecutive packets. Even so, a small number of neglected lost packets represents negligible failure in terms of first-alert monitoring. If the monitoring policy requires small numbers of packet loss to be accurately detected, or if this kind of overlooking is frequently observed in monitoring, combining delay variations can effectively reduces overlooking degradation in performance. For example, delay variation resulting from the delay of the first packet arises in the second case and are easily detected.

## 6. Applying Proposed Method to Real Monitoring Operations

A fixed packet length and an ideal IPG are needed in our method to distinguish VoIP packets and to compute the metrics of degradation a priori. These should be analyzed as explained in Sect. 3 before monitoring. If the ISP distributes TAs to customers, the packetization rate and codec with which the packet length and the IPG are derived can be known a priori.

Our flow registration algorithm (described in Sect. 4.2) can be applied to a high-speed link, e.g., 10 Gbit/s, for the following reasons. We found that UDP packets accounted for only 5.3% of packets in the Abilene and 5.2% of those in ISP traffic traces. 200-byte packets represented 7.9% of the total UDP packets for Abilene and 60.9% of the total of ISP. Hence, 200-byte UDP packets accounted for less than 4% of the total even in the ISP traffic traces. This ratio is sufficiently low for the algorithm to monitor VoIP flows even in a 10-Gbit/s link. Even if VoIP flows become more prevalent, e.g., half of all flows, this should be easy to implement in hardware to scale for larger numbers of VoIP flows because of the proposed algorithm's simplicity.

A maximum of 100% of VoIP flows may need to be monitored to estimate the degraded-packet rate with sufficient accuracy using our method. Even in this unlikely situation, i.e., if a 10-Gbit/s link were only filled with VoIP flows (156-K flows), the recent LSI technology would enable flows to be monitored even with such large amounts of flows.

Even if packet sampling is introduced before VoIP packets are distinguished to reduce the number of packets that must be treated, the method can still be used to estimate delay variations. In this case, un-sampled packets are treated as consecutive lost packets and $k_j^i$ is frequently larger than one. Therefore, the degraded-packet rate becomes much higher than the actual packet-loss rate. However, delay variations can still be estimated.

The comparison of metrics obtained from two monitoring points can be used to estimate the level of degradation between them. From analyzing the distribution for delay variations and the rate of consecutive packet loss, the flows monitored at the two points should be synchronized for the degraded-packet rate, since it varies greatly over flows. However, estimated delay variations for IPG deviations are expected to have similar values even if different flows are monitored between two monitoring points. Developing an actual method to synchronize monitored VoIP flows to obtain the degraded-packet rate remains as future work.

## 7. Conclusions

We proposed a lightweight method of extracting VoIP flows from Internet traffic by using knowledge about the specific features of VoIP flows, e.g., fixed packet sizes and long durations. It can distinguish VoIP flows in a high-speed links, e.g., 10 Gbit/s. We also proposed a method of estimating the delay variation and the rate of consecutive packet loss using a characteristic of VoIP flows, i.e., IPG is constant for VoIP flows.

Our evaluation using simulation with actual traffic showed that delay variation (IPG variance) could be accurately estimated by monitoring only a few percent of all flows. In contrast, estimating of the packet loss rate with error of around 0.2% requires that large numbers (more than 20%) of flows be monitored. The packet loss rate can also be estimated with the proposed method using single-point monitoring. It is therefore feasible to estimate VoIP performance by monitoring IPGs for a few flows, whereby ISPs will be able to assess whether or not VoIP service quality is being seriously degraded.

Further studies need to be done on assigning appropriate value to the receiving-TA buffer factor, which influences the detection of packet loss, reflecting the efficiency of monitoring policy. We also need details on how to implement the proposed method, e.g., how to harmonize the data obtained between ingress and egress points, and how to enhance it to run applications using variable bit-rate codec.

## Acknowledgments

This work was supported in part by the Japan Society for the Promotion of Science, with a Grant-in-Aid for Scientific Research (S) (No. 18100001).

### References

[1] "Information and communications in Japan," white paper, Ministry of Internal Affairs and Communications, 2006. http://www.johotsusintokei.soumu.go.jp/whitepaper/eng/WP2006/2006-index.html

[2] "Multi router traffic grapher," http://www.mrtg.org/

[3] "NetFlow service and applications," white paper, Cisco Systems, Inc., 2002. http://www.cisco.com/warp/public/cc/pd/iosw/ioft/neflct/tech/napps_wp.htm

[4] P. Phaal, S. Panchen, and N. McKee, "InMon corporation's sFlow: A method for monitoring traffic in switched and routed networks," RFC 3176, Internet Engineering Task Force, Sept. 2001.

[5] M. Grant and S. Tenissen, "Voice quality monitoring for VoIP networks," white paper. http://www.calyptech.com/pdf/CAL-000006-WP-01.pdf

[6] "Single-ended method for objective speech quality assessment in narrowband telephony applications," Telecommunication Standardization Sector of International Telecommunication Union Recommendation P.563, International Telecommunication Union, 2004.

[7] "The E-model, a computational model for use in transmission planning," Telecommunication Standardization Sector of International Telecommunication Union Recommendation G. 107, International Telecommunication Union, 2004.

[8] L. Deri, "Open source VoIP traffic monitoring," SANE 2006, Technical Report R3, May 2006. http://www.sane.nl/sane2006/program/final-papers/R3.pdf

[9] L. Deri, "nProbe: An open source NetFlow probe for gigabit networks," Proc. TERENA Networking Conference 2003, May 2003. http://luca.ntop.org/nProbe.pdf

[10] L. Deri, R. Carbone, and S. Suin, "Monitoring networks using ntop," Proc. IM 2001, pp.199–212, May 2001.

[11] "Abilene," http://abilene.internet2.edu/

[12] T. Kitamura, T. Shizuno, T. Okabe, and H. Tani, "Traffic identification for dependable VoIP," NEC Technical J., vol.1, no.3, pp.17–20, 2006.

[13] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, "RTP: A transport protocol for real-time applications," RFC 1889, Internet Engineering Task Force, Jan. 1996.

[14] "Pulse code modulation (PCM) of voice frequencies," Telecommunication Standardization Sector of International Telecommunication Union Recommendation G. 711, International Telecommunication Union, 1998.

[15] "Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear-prediction," Telecommunication Standardization Sector of International Telecommunication Union Recommendation G. 729, International Telecommunication Union, 1996.

[16] "End-to-end quality of service in TIPON systems; Part7: Design guide for elements of a TIPHON connection from an end-to-end speech transmission performance point of view," ETSI TS 101 329-7, European Telecommunications Standards Institute, 2002.

**Yoshinori Kitatsuji** received his B.E. and M.E degrees from Osaka University, Japan in 1995 and 1997, respectively, and received his D.E. degree from Kyushu Institute of Technology, Japan in 2007. In 1997, he joined the KDD Co., Ltd. He was a research fellow of Telecommunications Advancement Organization of Japan from 2001 to 2003, and then was an expert researcher in Kyushu Research Center, National Institute of Information and Communications Technology. He is currently a research engineer of Mobile Network Laboratory in KDDI R&D Laboratories Inc. His research interests include performance monitoring, and traffic engineering of computer communication networks.

**Satoshi Katsuno** received the B.S. degree and the M.S. degree in electric engineering from the University of Tokyo, Japan in 1989 and 1991, respectively. He joined KDD in 1991. He worked at the Tokyo Research & Operation Center of Telecommunication Advancement Organization (TAO) from 2001 to 2004. Since 2004, he has been at KDDI R&D Laboratories Inc. He has been engaged in research on network quality measurement.

**Katsuyuki Yamazaki** received B.E. and D.E degrees from the University of Electro-Communications and Kyushu Institute of Technology in '80 and '01, respectively. At KDDI Co. Ltd., he had been engaged in R&D and international standardization of ISDN and S.S. No.7, ATM networks, L2 networks, Internet and ubiquitous networking, and was responsible for R&D strategy of KDDI Labs Inc. He is currently a Professor of Nagaoka University of Technology.

**Masato Tsuru** received B.E. and M.E. degrees from Kyoto University, Japan in 1983 and 1985, respectively, and then received his D.E. degree from Kyushu Institute of Technology, Japan in 2002. He worked at Oki Electric Industry Co., Ltd., Information Science Center, Nagasaki University, Japan Telecom Information Service Co., Ltd., and Telecommunications Advancement Organization of Japan. In 2003, he moved to the Department of Computer Science and Electronics, Faculty of Computer Science and Systems Engineering, Kyushu Institute of Technology as an Associate Professor, and then has been a Professor in the same department since April 2006. His research interests include performance measurement, modeling, and management of computer communication networks. He is a member of the IPSJ, JSSST and ACM.

**Yuji Oie** received B.E., M.E. and D.E. degrees from Kyoto University, Kyoto, Japan in 1978, 1980 and 1987, respectively. From 1980 to 1983, he worked at Nippon Denso Company Ltd., Kariya. From 1983 to 1990, he was with the Department of Electrical Engineering, Sasebo College of Technology, Sasebo. From 1990 to 1995, he was an Associate Professor in the Department of Computer Science and Electronics, Faculty of Computer Science and Systems Engineering, Kyushu Institute of Technology, Iizuka. From 1995 to 1997, he was a Professor in the Information Technology Center, Nara Institute of Science and Technology. Since April 1997, he has been a Professor in the Department of Computer Science and Electronics, Faculty of Computer Science and Systems Engineering, Kyushu Institute of Technology. His research interests include performance evaluation of computer communication networks, high speed networks, and queuing systems. He is a fellow of the IPSJ and a member of the IEEE.