Effect of Packetization on VoIP Performance

Boonchai Ngamwongwattana
National Electronics and Computer Technology Center (NECTEC)
112 Phahol Yothin Rd., Klong Luang, Pathumthani 12120, THAILAND
Email: boonchai.ngamwongwattana@nectec.or.th

Abstract—**The fundamental characteristics of VoIP are constant rate, delay sensitivity, and loss tolerance. VoIP over packet-switched networks including the Internet poses problems because the network service is not guaranteed to meet such requirements, i.e. the available bandwidth as well as the delay and loss bounds. Adaptive rate VoIP is a solution that can mitigate the problem. Adaptive rate VoIP has the ability to adapt its transmission rate to match the available network bandwidth. This helps to reduce or avoid network congestion, which in turn minimizes delay and loss. To implement adaptive rate VoIP, the VoIP source must be able to send packets at different rates. Adaptive multi-rate speech coders are commonly used. However, their voice quality (e.g. MOS) varies depending on the bitrate. In this paper, we propose an alternative of using packetization as a means for rate adaptation while using a constant bitrate coder. We explore how packetization can vary network bandwidth requirement. We then study the effect of packetization on VoIP performance. The simulation study shows an interesting result. Using an optimal packetization can help to improve VoIP performance. At the same time, the amount of voice traffic plays an important role to determine the performance improvement. This study also demonstrates that feasibility of packetization-based adaptive rate VoIP.**

Index Terms—**Voice over IP (VoIP), Adaptive Rate VoIP, Packetization, Performance**

I. Introduction

As is well known, the Internet service is not guaranteed. It is possible at any time that a network overload causes high delay and packet loss. TCP has a congestion control mechanism embedded to back-off in the event of packet loss, ensuring that network congestion can be resolved. Real-time applications and VoIP in particular, on the other hand, usually send packets at a constant rate with no control mechanism. The inherent problem is that they cannot react to network congestion, causing problems to the performance. Adaptive rate VoIP is expected to a solution to mitigate the problem. A barrier for adaptive rate VoIP is speech coder. Variable bitrate speech coders were virtually non-existent in the past. Speech coders are generally model-based, sending parameters which represent the model independent of network conditions. Some researchers proposed the use of banks of speech coders; each with different bitrates, and switching between them to perform adaptive rate control [1]. This approach has some drawbacks. One is the problem of implementing many speech coders in the same platform. Also, the transition from one coder to another might not be transparent to the user and might cause some distraction. Until recently, variable bitrate speech coders have been developed. A well-known example is the GSM Adaptive Multi-Rate (AMR) speech coder [2]. The AMR coder supports 8 different bitrates, which the voice quality (e.g. MOS) varies depending on the bitrate. The higher the bitrate, the better the voice quality [3].

In this paper, we propose an alternative of using packetization as a means for rate adaptation for adaptive rate VoIP. An advantage is that it works with any constant bitrate speech coder. In addition, rate adaptation can be achieved without having an impact on the voice quality. Hence, rate adaptation is transparent and does not cause distraction to the user. Given a constant bitrate speech coder, we explore how packetization can vary network bandwidth requirement. We then study the effect of packetization on VoIP performance.

II. Packetization and Bandwidth Requirement

One decision faced by VoIP users is how many sample frames from the speech coder to include in the packet payload, or called packetization. This is an important issue because packetization determines the payload size as well as the network bandwidth requirement. Since VoIP is delay-sensitive, a small payload size is needed so that it does not cause too much delay from collecting the sample frames. A typical VoIP packet requires at least 40 bytes of overhead (20 bytes of the IP header, 8 bytes of the UDP header, and 12 bytes of the RTP header). The overhead of the data link layer is usually not considered because it varies as the packet travels across different physical networks. The size of the packet overhead is usually larger than the payload. Thus, a large percentage of bandwidth is used for the transport of overhead bytes. Table I shows important characteristics of well-known speech coders that are needed for bandwidth requirement calculation. Sample frame delay refers to the time interval in which the coder samples voice signal, encodes it, and outputs a digitized voice frame. Sample frame size is the size in bits of

TABLE I
Speech Coder Characteristics Used for Bandwidth Calculation

Speech Coder	Effective Voice Bandwidth (Kbps)	Sample Frame Delay (ms)	Sample Frame Size (bits)
G.711 PCM	64	0.125	8
G.726 ADPCM	32	0.125	4
G.729 CS-ACELP	8	10	80

978-1-4244-2101-5/08/$25.00 ©2008 IEEE

the digitized voice frame, which is the smallest data unit that can be placed in the packet payload.

To explore the relationship between packetization and bandwidth requirements, we write the following equations. Let

T	Sample frame delay (msec)
F	Sample frame size (bits)
n	Number of sample frames in the payload
H	Header size of the voice packet (bits)

$$\text{The effective voice bandwidth} = \frac{F}{T}\,\text{Kbps} \tag{1}$$

$$\text{The overhead bandwidth} = \frac{H}{nT}\,\text{Kbps} \tag{2}$$

$$\text{The network bandwidth} = \frac{H+nF}{nT}\,\text{Kbps} \tag{3}$$

The effective voice bandwidth in (1) is the output bitrate of the speech coder. Equation (2) is the amount of bandwidth consumed by the packet overhead. The total network bandwidth required by a VoIP session is the sum of (1) and (2), which gives (3). Notice from the equations that placing more sample frames into the payload helps to reduce overhead bandwidth as well as network bandwidth.

Fig. 1 is a composite plot of the above equations, using parameters from the ADPCM coder. The lower horizontal scale is the number of sample frames in the payload. The upper horizontal scale shows the corresponding packetization delay of the lower scale, which is the product of sample frame delay and the number of sample frames in the payload. The gap between the network bandwidth and the effective voice bandwidth is the overhead bandwidth. As seen in the figure, the overhead and network bandwidth exhibits an exponential decrease as a function of packetization. Small packetization results in a low payload-to-overhead ratio. That is, in addition to the effective voice bandwidth, a large percentage of bandwidth is required for the transport of packet overhead. This causes extremely large network bandwidth requirement. On the other hand, large packetization helps to increase the payload-to-overhead ratio, which helps to reduce the overhead as well as the network bandwidth requirement. Large packetization, however, also causes excessive amount of time being used for waiting for many sample frames, or called packetization delay. Since VoIP is delay-sensitive, this impacts the time remaining to meet the end-to-end delay budget for acceptable voice quality. A typical solution to the trade-off of packetization is to choose a reasonable value that poses a moderate network bandwidth requirement as well as a moderate packetization delay. From Fig. 1, 160-frame packetization is an example.

Since network condition varies over time, a VoIP system using a pre-determined packetization would not be able to match its transmission rate to the available bandwidth. As illustrated in Fig. 1, we believe that packetization can be used as a means for rate adaptation for adaptive rate VoIP. By varying packetization (also varying payload size), a control mechanism will be able to choose an optimal transmission rate to maximize the performance. When the network is lightly loaded, using a small payload allows minimal packetization delay. As the network load increases, using a larger payload can help to reduce the transmission rate accordingly, at the expense of larger packetization delay. Note that, as seen in the figure, too large packetization does not give much benefit. The network bandwidth requirement slightly decreases, while packetization delay increases significantly. Similarly, too small packetization causes extremely large network bandwidth requirement, hence not beneficial as well. Therefore, a feasible range of packetization should only be used.

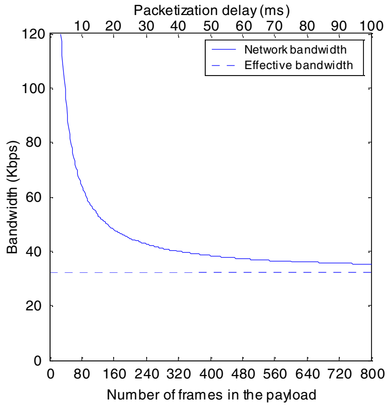

Fig. 1. Relationship between packetization and bandwidth requirements, using parameters from the ADPCM coder.

Adaptive rate VoIP based on packetization has an advantage that it can use any constant bitrate speech coder. Adapting the transmission can be achieved without having an impact on the original voice quality. Hence, the perceived quality would be more transparent and less distraction to the user. A side effect of varying packetization is the added packetization delay, which is manageable to ensure the end-to-end delay within an acceptable range. Compared to variable bitrate speech coders, although they do not affect packetization delay, it affects the voice quality. When the coder lowers its bitrate, the output voice quality is degraded, which can easily cause distraction to the user.

III. Simulation and Results

A. Simulation Setup

We conduct a simulation study using the Network Simulator 2 or ns-2 [4]. The network topology for the simulations is shown in Fig. 2. All nodes implement FIFO scheduling and drop-tail queuing. The link between node 0 and 1 has capacity of 10 Mbps with propagation delay of 35 milliseconds. The cross traffic over this link creates load around 60 percent on average. The link between node 1 and 2 has limited capacity so as to create a bottleneck and node 1 is the bottleneck point. The link capacity varies as a factor of the simulations, with propagation delay of 5 milliseconds. The cross traffic on each link is generated from nine Pareto sources

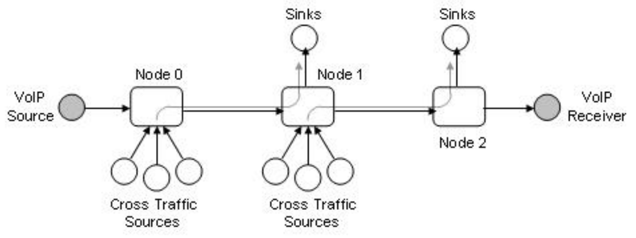

Fig. 2. Network topology for the simulation.

with α of 1.5, i.e. the inter-arrival times have infinite variance. The aggregation of many Pareto sources with α less than 2 has been shown to produce Long Range Dependent (LRD) traffic [5]. Measurement studies have shown that packet size distribution in the Internet is centered on three values [6, 7]. Specifically, about 60% of the packets are 40 bytes, 25% are 550 bytes, and 15% are 1500 bytes. In the simulation, packet sizes of the cross traffic are distributed following these findings. Note that, in terms of load distribution, about 7% of the packets are 40 bytes, 35% are 550 bytes, and 58% are 1500 bytes. In each simulation, the VoIP source sends packets at a constant rate, through all nodes, to the receiver. The voice traffic lasts for 120 seconds. The source is assumed using the ADPCM coder, with the voice bandwidth of 32 Kbps.

Packetization and the level of cross traffic load over the bottleneck link are the key factors in the simulation study. We consider the feasible range of packetization from 80 to 240 bytes of payload, or from 10 to 30 milliseconds of packetization delay. Another important factor is the percentage of voice traffic over the bottleneck link. Specifically, we define impact factor as the effective voice bandwidth divided by the bottleneck link capacity.

In each simulation, we make measurements of packet loss rate as well as one-way network delay of voice packets arriving at the receiver. Note that the network delay is only associated with the packet level. To evaluate the performance of VoIP, the measurement must be in the sample frame level. That is, packetization delay must be included. Here, we define one-way end-to-end delay as the latency of a sample frame from which it is outputted from the coder to which it is received by the decoder. The end-to-end delay can be found by which it is the sum of the measured network delay and the packetization delay.

Descriptive statistics are used to characterize performance among different factors, as well as to make comparative evaluation. We use the 90th percentile of end-to-end delay, instead of the commonly used mean end-to-end delay. In VoIP, packets arriving on time must wait for late packets so that all the packets can be played out smoothly. The 90th percentile of end-to-end delay is a better metric than the mean because it estimates the actual perceived delay that the user would experience. In other words, the 90th percentile of end-to-end delay virtually accounts for an estimate of the jitter buffer delay.

B. Effect of Packetization on VoIP Performance

The perceived quality of VoIP is generally determined by the delay and loss performance. However, making a comparative evaluation among pairs of delay and loss measurements is not a simple task. For instance, which has better perceived quality between (150-ms delay, 4-percent loss) and (250-ms delay, 2-percent loss)? This usually requires perceived quality assessment. In addition, delay jitter is another key factor that can significantly affect the perceived quality. The subjective methods are known to be much suitable for VoIP quality assessment, but they require a great deal of resources. In this simple study, we attempt to use an objective method to evaluate the VoIP performance. To avoid the evaluation problems, we primarily pay attention to the levels of cross traffic load that result in relatively low packet loss. This allows us to focus on evaluating the performance using the end-to-end delay. Fig. 3, 4, 5, and 6 are the simulation results showing the effect of packetization when the bottleneck link capacity is 128, 256, 512, and 768 Kbps, respectively.

The figures demonstrate the effect of packetization. We choose Fig. 4 to explain the findings. The convex curves in the figure show the important inherent trade-off of packetization. Recall that end-to-end delay is the sum of packetization delay and network delay (primarily due to queuing delay). Ideally, small packetization is desirable so as to minimize packetization delay. However, this results in a huge network bandwidth requirement. If the network cannot afford to provide such an available bandwidth, it causes an increase in queuing delay. The large queuing delay can significantly outweigh the saving in the packetization delay, causing large end-to-end delay as a result. This happens to the left side of the convex point. On the other hand, large packetization causes an additional packetization delay and is usually not desirable. However, this helps to reduce the network bandwidth requirement. It is more likely that the network has sufficient available bandwidth to a smaller requirement; hence, does not cause an increase in queuing delay. To the right side of the convex point, the end-to-end delay is more affected by packetization delay, rather than queuing delay. A key conclusion from this study is that, given a network condition, there is an optimal packetization that allows the network bandwidth requirement to match the available network bandwidth. This results in the minimal end-to-end delay.

From Fig. 4, the plot of the 40-percent cross traffic load is a straight line, instead of a convex curve. This happens because the network load is so low. The available network bandwidth is more than sufficient to support the bandwidth required by the VoIP session, even at 10-ms packetization. The end-to-end delay is, hence, only affected by packetization delay. In this case, the convex curve will appear when the packetization is far less than 10 milliseconds.

Take the network load factor into consideration, from Fig. 4, the overall end-to-end delay increases more dramatically when the network is heavily loaded. This is primarily due to the queuing delay. As the network load increases, available network bandwidth decreases. The optimal packetization needs to be larger in order to lower the network bandwidth

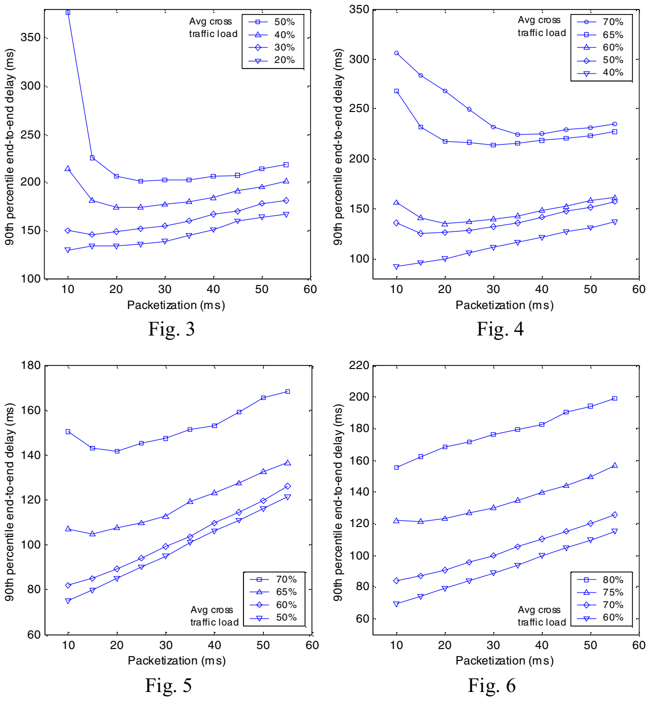

Fig. 3

Fig. 4

Fig. 5

Fig. 6

Fig. 3, 4, 5, and 6. Effect of packetization on end-to-end delay for the bottleneck link of 128, 256, 512, and 768 Kbps, respectively.

requirement. Consider the 65-percent and 70-percent traffic load curves, in which the network is heavily loaded, optimizing packetization helps significantly to minimize the end-to-end delay. This result suggests that using a pre-determined packetization is not an effective implementation because the optimal packetization depends on the network load. Adaptive rate VoIP based on packetization would be able to take advantage of this finding.

The impact factor plays a role in the effectiveness of optimizing packetization. From Fig. 3, 4, 5, and 6, the impact factor is at 25, 12.5, 6.25, and 4.16 percent, respectively. Except for the impact factor of 4.16 percent, all other figures show the convex curves. As the impact factor gets smaller, as seen from the figures, the benefit of optimizing packetization diminishes. A small impact factor means that the VoIP traffic is only a small fraction of the bottleneck link capacity. Thus, the change in the network bandwidth requirement cannot make a significant impact on the overall traffic load. In this case, the performance improvement would be tiny to be noticeable. Nonetheless, the results suggest that optimizing packetization starts to yield a benefit when the impact factor is around 5 percent, which is a fairly small percentage.

From Fig. 6, the network has a relatively large capacity bottleneck. Any packetization gives no difference to the network delay. Large packetization worsens the end-to-end delay because it is affected by the increase in packetization delay. In this environment, the VoIP traffic is a minority and the performance is largely caused by the overall traffic load. Optimizing packetization still has a benefit though. The strategy in this case is to packetize as small as possible to ensure the minimal end-to-end delay. A typical VoIP cannot achieve the same minimal end-to-end delay because it usually uses a moderately large packetization.

IV. Conclusion

In this paper, we propose an alternative of using packetization as a means for rate adaptation for adaptive rate VoIP. The primary goal of this work is to study the effect of packetization on VoIP performance. We have shown the inherent trade-off of packetization that is critical to minimizing the end-to-end delay performance. The simulation results show the three factors that affect VoIP performance: packetization, cross traffic load, and impact factor. Optimizing packetization can help to match the network bandwidth requirement to the available network bandwidth. Under different levels of network load, there is an optimal packetization that allows minimal end-to-end delay. The results and findings also demonstrate the feasibility of adaptive rate VoIP based on packetization.

References

[1] J.-C. Bolot, and A. Vega-Garcia, "Control mechanisms for packet audio in the internet," in *Proc. IEEE INFOCOM,* San Francisco, USA, Mar. 1996.

[2] ETSI EN 301 704 V7.2.1 (2000-04), Digital cellular telecommunications system (Phase 2+); Adaptive Multi-Rate (AMR) speech transcoding (GSM 06.90 version 7.2.1 Release 1998).

[3] F. Beritelli, G. Ruggeri, G. Schembra, "TCP-Friendly Transmission of Voice over IP," in *Proc. IEEE ICC*, Apr. 2002.

[4] The Network Simulator – ns-2, http://www.isi.edu/nsnam/ns/.

[5] M. S. Taqqu, W.Willinger, and R. Sherman, "Proof of a Fundamental Result in Self-Similar Traffic Modeling," *ACM Computer Communications Review*, pp. 5 – 23, Apr. 1997.

[6] K. Claffy, G. J. Miller, and K. Thompson, "The nature of the beast: recent traffic measurement from an Internet backbone," in *Proc. INET '98*, Geneva, Switzerland, Jul. 1998.

[7] K. Thompson, G. J. Miller, and R. Wilder, "Wide-Area Internet Traffic Patterns and Characteristics," *IEEE Network*, pp. 10 – 23, Nov. 1997.