Adaptive Voice Smoother with Optimal Playback Delay for New Generation VoIP Services

Shyh-Fang Huang [1], Eric Hsiao-Kuang Wu [2], and Pao-Chi Chang [3]

[1] Department of Electrical Engineering, National Central University, Taiwan,
hsf@vaplab.ee.ncu.edu.tw

[2] Department of Computer Science & Information Engineering, National Central University, Taiwan,
hsiao@csie.ncu.edu.tw

[3] Department of Communications Engineering, National Central University, Taiwan,
pcchang@ee.ncu.edu.tw

Abstract. Perceived voice quality is a key metric in VoIP applications. The quality is mainly affected by IP network impairments such as delay, jitter and packet loss. Playout buffer at the receiving end can be used to compensate for the effects of jitter based on a tradeoff between delay and loss. Adaptive smoothing algorithms are capable of adjusting dynamically the smoothing time based on the network parameters to improve voice quality. In this article, we introduce an efficient and easy perceived quality method for buffer optimization to archive the best voice quality. This work formulates an online loss model which incorporates buffer sizes and applies the Lagrange multiplier approach to optimize the delay-loss problem. Distinct from the other optimal smoothers, the proposed optimal smoother is suitable for any codec and carries the lowest complexity. Simulation experiments validate that the proposed adaptive smoother archives significant improvement in the voice quality.

1 Introduction

The rapid progress of the development of IP-base network has enabled numerous applications that deliver not only traditional data but also multimedia information in real time. The next generation network, like an ALL-IP network, is a future trend to integrate all heterogeneous wired and wireless networks and provide seamless world-wide mobility. In an All-IP network, one revolution of the new generation Internet applications will realize VoIP services that people can talk freely around through the mobile-phones, the desktops and VoIP telephones at any time and place. Unfortunately, the IP-based networks do not guarantee the available bandwidth and assure the constant delay jitters (i.e., the delay variance) for real time applications. In other words, individual transmission delays for a given flow of packets in a network may be continuing to change caused by varying traffic load and differing routing paths due to congestions, so that the packet network delays for a continuous series of intervals (i.e.

talkspursts) at the receiver may not be the same (i.e. constant) as the sender. In addition, a packet delay may introduce by the signal hand-out or the difference of bandwidth transportation in wireless/fixed networks.

For delay sensitive applications, a dominating portion of packet losses might be likely due to delay constraint. A late packet that arrives after a delay threshold determined by playback time is treated as a lost packet. A tight delay threshold not only degrades the quality of playback but also reduces the effective bandwidth because a large fraction of delivered packets are dropped. In fact, delay and loss are normally not independent of each other. In order to reduce the loss impact, a number of applications utilize an adaptive smoothing technique in which buffers are adopted to reduce the quality damage caused by loss packets. However, a large buffer will introduce excessive end-to-end delay and deteriorate the multimedia quality in interactive real-time applications. Therefore, a tradeoff is required between increased packet loss and buffer delay to achieve satisfactory results for playout buffer algorithms.

In the past, the works on the degradation of the voice quality consider the effect of packet loss, but not that of packet delay. Within literature on predicting delays, the use of Pareto distribution in [1] is of computing the distribution parameters and rebuilding the new distribution to predict the next packet delay, and the use of neural network models to learn traffic behaviors [2]. The use of Pareto distribution or a neural network model requires relatively high complexity or a long learning period. Therefore, we consider the smoothers [3]-[9] which employ statistical network parameters related with the voice characteristic, i.e. loss, delay and talk-spurt that have significant influence to the voice quality. They detect delay spike in traffic and quickly calculate the required buffer size to keep the quality as good as possible.

For perceptual-based buffer optimization schemes for VoIP, voice quality is used as the key metric because it provides a direct link to user perceived QoS. However, it requires an efficient and accurate objective way to optimize perceived voice quality. In this paper, we propose a new delay-loss smoother that employs the Lagrange multiplier method to optimize the voice quality by balancing the delay and the loss. Lagrange multiplier method is often used to optimize the trade off problems. The contributions of this paper are two-fold: (i) A new method is for optimizing voice quality for VoIP and is easily applied to new codecs. (ii) Different from the other optimal smoothers, our optimal smoother has the lowest complexity with $O(n)$. The remainder of the paper is structured as follows. Section 2 reviews the previous work. Section 3 proposes the novel adaptive smoother. Section 4 shows the simulation results in smoothers. Finally, conclusions are provided.

2 Related Work

The SD algorithm has been studied by many researchers [3]-[9]. A delay spike is defined as a sudden and significant increase of network delay in a short period often less than one round-trip. This algorithm adjusts the smoothing size, i.e. playback delay, at the beginning of each talk-spurt. The results of this algorithm are therefore compared to the results obtained herein.

The SD Algorithm in [3] estimates the playout time p_i of the first packet in a talk-spurt from the mean network delay d_i and the variance v_i for packet i as

$$p_i = t_i + d_i + \gamma v_i \tag{1}$$

where t_i represents the time at which packet i is generated at the sending host and γ is a constant factor used to set the playout time to be "far enough" beyond the delay estimate such that only a small fraction of the arriving packets could be lost due to late arrival . The value of $\gamma = 4$ is used in simulations [3]. The estimates are recomputed each time a packet arrives, but only applied when a new talk-spurt is initiated.

The mean network delay d_i and variance v_i are calculated based on a linear recursive filter characterized by the factors α and β.

$$\begin{cases} If\ n_i > d_{i\text{-}1} \Rightarrow \begin{cases} d_i = \beta d_{i-1} + (1-\beta) n_i \\ v_i = \beta v_{i-1} + (1-\beta)\left| d_{i-1} - n_i \right| \end{cases} & (SPIKE\_MODE) \\ If\ n_i \le d_{i\text{-}1} \Rightarrow \begin{cases} d_i = \alpha d_{i-1} + (1-\alpha) n_i \\ v_i = \alpha v_{i-1} + (1-\alpha)\left| d_{i-1} - n_i \right| \end{cases} \end{cases} \tag{2}$$

where n_i is the end-to-end delay introduced by the network and typical values of α and β are 0.998002 and 0.75 [3], respectively.

The decision to select α or β is based on the current delay condition. The condition $n_i > d_{i-1}$ represents network congestion (*SPIKE_MODE*) and the weight β is used to emphasize the current network delay. On the other hand, $n_i \le d_{i-1}$ represents network traffic is stable, and α is used to emphasize the long-term average.

In estimating the delay and variance, the SD Algorithm uses only two values α and β that are simple but may not be adequate, particularly when the traffic is unstable. For example, an under-estimated problem is when a network becomes spiked, but the delay n_i is just below the d_{i-1}, the SD Algorithm will judge the network to be stable and will not enter the *SIPKE_MODE*.

3 Optimal Smoother with Delay-Loss Trade off

The proposed optimal smoother is derived using the Lagrangian method to trade off the delay and loss. This method involves, first, building the traffic delay model and the loss model. Second, a Lagrangian cost function Q is defined using this delay and the loss models. Third, the Lagrangian cost function Q is minimized and thus the delay and loss optimized solution is obtained.

3.1 Traffic Delay and Loss Models

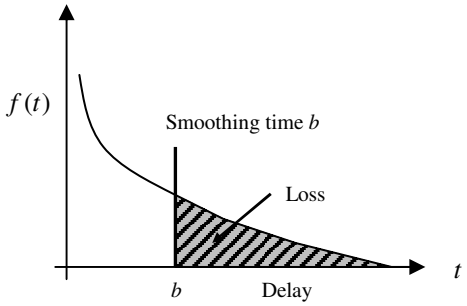


Fig. 1. The relation of smoothing delay and loss

For perceived buffer design, it is critical to understand the delay distribution modeling as it is directly related to buffer loss. The characteristics of packet transmission delay over Internet can be represented by statistical models which follow Exponential distribution for Internet packets (for a UDP traffic) has been shown to consistent with an Exponential distribution [10]. In order to derive an online loss model, the packet end-to-end delay is assumed as an exponential distribution with parameter $1/\mu$ at the receiving end for low complexity and easy implementation. The CDF of the delay distribution $F(t)$ can also be represented by [11]

$$F(t)=1-e^{tu^{-1}} \tag{3}$$

and the PDF of the delay distribution $f(t)$ is

$$f(t)=\frac{dF(t)}{dt}=\mu^{-1}e^{-t\mu^{-1}} \tag{4}$$

In a real-time application, a packet loss that is solely caused by extra delay can be derived from the delay model $f(t)$. Figure 1 plots the delay function $f(t)$, which shows that when the packet delay exceeds the smoothing time; the delayed packet is regarded as a lost packet. The loss function $l(b)$ can be derived from Fig. 1 as

$$l(b)=\int_b^{\infty}f(t)dt=\left(-e^{-\mu^{-1}t}\right)\Big|_b^{\infty}=-e^{-\infty}+e^{-\mu^{-1}b}=e^{-\mu^{-1}b} \tag{5}$$

From Eqs. (4) and (5), we obtain the delay and loss functions that will be used in Lagrangian cost function.

3.2 Optimal Delay-Loss Adaptive Smoother

To express the corresponding quality for a given voice connection, a Lagrangian cost function Q is defined based on the delay b and the loss model $l(b)$

$$Q(b) = b + K \cdot l(b) \tag{6}$$

where $Q(b)$ represents the negative effect on voice quality, i.e., minimizing Q yields the best voice quality. K is a Lagrange multiplier where the loss becomes more significant as K increases. The K value has significant influence on the optimization process. We will discuss the valid range of the value in this section and the suggested value in the next section.

Here, once a smoothing time b is specified, the loss $l(b) = e^{-\mu^{-1}b}$ can be calculated from Eq. (5). The Lagrangian cost function in Eq. (6) yields

$$Q(b) = b + K \cdot e^{-\mu^{-1}b} \tag{7}$$

The differential equation dQ/db is assigned to zero that minimizes Q to yield the smoothing time b,

$$b = \mu \ln\left(K\mu^{-1}\right) \tag{8}$$

where b is the best smoothing time for balancing the delay and the loss. Afterwards, the smoother can provide best quality, considering both the delay and the loss effects, based on the calculated smoothing time b.

The calculated smoothing time b is a function of K and μ. μ denotes a IP-base network delay parameter (end-to-end delay) and can be measured at the receiver, but K is given by users or applications. The calculated smoothing time b must be within an allowable range to ensure that the end-to-end delay is acceptable. Here, D_{max} is defined as the maximum acceptable end-to-end delay and the calculated smoothing time b must be between 0 and D_{max}

$$0 \le b = \mu \ln\left(K\mu^{-1}\right) \le D_{max} \tag{9}$$

Accordingly, the permissible range of valid K in the Lagrange multiplier Q function in Eq. (8) is

$$\mu \le K \le e^{D_{\max} * \mu^{-1}} * \mu \tag{10}$$

3.3 Suggestion of *K* Parameter

In this section the relationship between the voice quality and loss is further studied. Based on the previous section discussions, we know K parameter is tightly related with voice quality. In other words, for a given MOS (Mean Opinion Score) of

speech quality, the allowable range of K can further be restricted. Many studies revealed the difficulty of determining the mathematical formula that relates the voice quality, delay, and loss. According to [12], the loss degrades the voice quality more remarkably than does the delay, so the quality-loss relationship is first emphasized [13] [14]. In these studies, an empirical Eq. (11) was obtained by experiments with many traffic patterns for predicting the voice MOS quality MOS_{pred} that might be degraded by the traffic loss ($loss$)

$$MOS_{pred} = MOS_{opt} - c * ln(loss + 1) \tag{11}$$

where MOS_{opt} is voice codec related, representing the optimum voice quality that the codec can achieve, c is a constant that is codec dependent, and $loss$ is a percentage ratio times 100. Following this approach, anyone can estimate a specific empirical rule with specified voice codecs and network environments. Equation (11) also implies that the network loss rate must be kept lower than or equal to the defined $loss$ to ensure the predicted MOS MOS_{pred} .

Equation (11) is rewritten to yield Eq. (12),

$$loss = 2^{\frac{MOS_{opt} - MOS_{pred}}{c}} - 1 \tag{12}$$

Notably, the $l(t)$ function is a percentage but loss is not. Therefore, $l(t)$ is multiplied by 100 to yield

$$loss = 2^{\frac{MOS_{opt} - MOS_{pred}}{c}} - 1 \geq l(t) = e^{-\mu^{-1} b} * 100 \geq 0 \tag{13}$$

From Eq. (13), the smoothing time b is

$$b \geq -ln\left(\frac{2^{\frac{MOS_{opt} - MOS_{pred}}{c}} - 1}{100} \right) * \mu \tag{14}$$

From Eqs. (8), (10) and (14), the suggested range for K is

$$K \geq max\left(\frac{100 * \mu}{(2^{\frac{MOS_{opt} - MOS_{pred}}{c}} - 1)}, \mu \right) \tag{15}$$

When K is assigned a value that is more than the threshold in Eq. (15), the design of the smoother is mainly dominated by the loss effect. For a given MOS, a suitable K can be suggested and an optimal buffer size can be determined.

4. Simulation

4.1 Simulation Configuration

A set of simulation experiments are performed to evaluate the effectiveness of the proposed adaptive smoothing scheme. The OPNET simulation tools are adopted to trace the voice traffic transported between two different LANs for a VoIP environment. Ninety personal computers with G.729 traffics are deployed in each LAN. The duration and frequency of the connection time of the personal computers follow Exponential distributions. Ten five-minute simulations were run to probe the backbone network delay patterns, which were used to trace the adaptive smoothers and compare the effects of the original with the adapted voice quality latter.

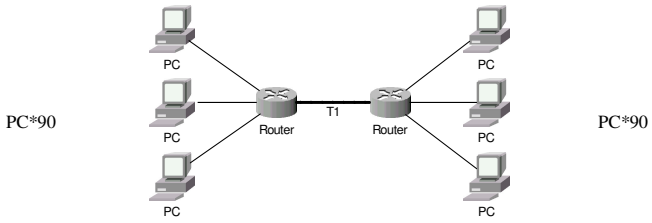


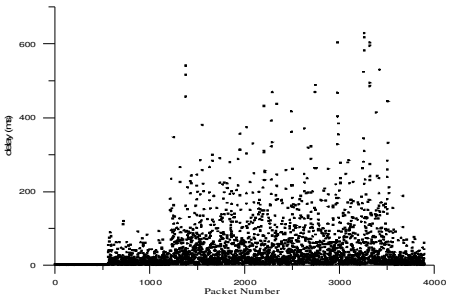
Fig. 2. The simulation environment of VoIP

Table 1. Simulation parameters

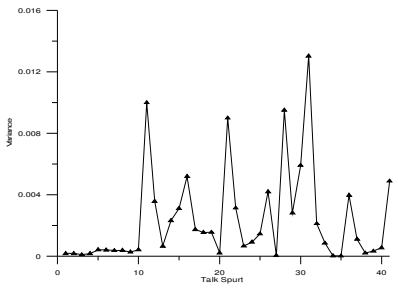
Attribute	Value
Numbers of PC in one LAN	90 PCs
Codec	G.729
Backbone	T1 (1.544 Mps)
LAN	100 Mbps
Propagation delay	Constant
Router buffer	Infinite
Packet size	50 bytes

Fig. 2 shows the typical network topology in which a T1 (1.544 Mbps) backbone connects two LANs, and 100 Mbps lines are connected within each LAN. The propagation delay of all links is assumed to be a constant value and will be ignored (the derivative value will be zero) in the optimization process. The buffer size of the bottlenecked router is assumed to be infinite since the performance comparison of adaptive smoothers will be affected by overdue packet loss (over the deadline) and not affected by the packet loss in router buffer. The network end-to-end delay of a G.729 packet with data frame size (10 bytes) and RTP/UDP/IP headers (40 bytes) is measured for ten five-minute simulations by employing the OPNET simulation network. Table 1 summarizes the simulation parameters. Figure 3(a) and 3(b) list one

of the end-to-end traffic delay patterns and the corresponding delay variances for VoIP traffic observed at a given receiver.



(a) The delay of traffic



(b) The variance of traffic

Fig. 3. VoIP traffic pattern

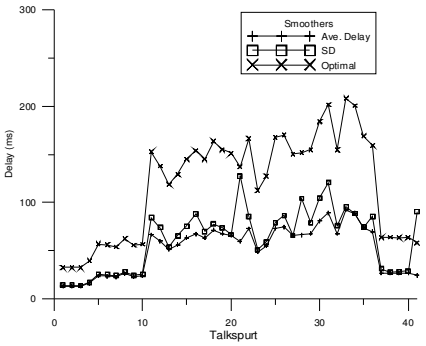


Fig. 4 The predicting time of smoothers

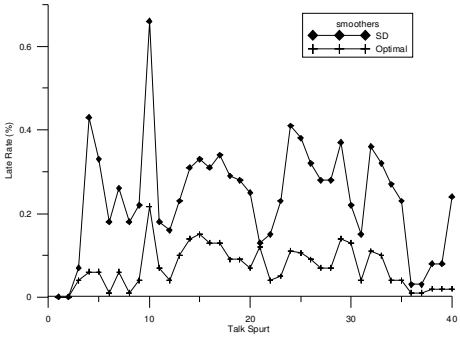


Fig. 5 The packet loss rate of smoothers

4.2 Predicted Smoothing Time and Loss Rate in Smoothers

In this section the accuracy of the predicted end-to-end delay time and loss rate among these smoothers are compared. The maximum acceptable delay D_{max} is set to 250 ms and the average delay is used to observe the traffic pattern in particular. In Fig. 4 and Fig. 5, we can observe that the predicted time of the SD smoother is very close to the mean delay and the loss rate is higher than optimal smoother. The SD smoother uses a large value of fixed β to deal with various traffic conditions and emphasize a long-term mean delay d_{i-1}, so the predicted delay will be close to the mean delay. A better choice for n_i is probably the maximum delay in the last talkspurt that can sufficiently represent the worst case of current network congestion and avoid an under-estimated delay.

4.3 Quality Measurement

The test sequence is sampled at 8 kHz, 23.44 seconds long, and includes English and Mandarin sentences spoken by male and female. Table 2 lists the mean delay, mean loss rate, and SSNR measured in a voice quality test with various smoothers. SSNR [15][16] is used as an evaluation tool because it correlates better with MOS and it is relatively simple to compute. Table 2 shows that the Optimal smoother performance achieves a high average SSNR and has the significant improvement in the voice quality over SD smoother, since the proposed optimal smoother truly optimizes with the delay and loss impairments.

The SSNR can only represent the loss impact, but hardly represent the delay impact. Therefore, a Lagrangian cost function is utilized to consider the delay and loss impacts to the quality degradation for various smoothers. In order to maintain the normal voice quality over the network, the predicted MOS, $MOS_{pred} = 3$ is required. According to [14] and G.729, c is set as 0.25 in formula (15) and the μ is set as the frame rate 10 ms for G.729 at the sender. The Lagrange multiplier value $K = 430$ is calculated from the formula (15). Table 3 shows that the optimal smoother has the lower Lagrangian cost value than SD smoother. Specifically, we can observe the optimal smoother has 23% improvement of the quality degradation on SD smoother.

Table 2. The voice quality test of smoothers

	Source	SD	Optimal
SSNR (dB)	8.17	5.67	7.51
Mean delay (ms)		89.22	112.46
Mean Loss Rate (%)		0.21	0.09

Table 3. The mean negative cost value of smoothers at high traffic load

Smoother	SD	Optimal
Lagrangian Cost (ms)	220.2166	170.2838

5. Conclusion

For new-generation VoIP services, a dynamic smoothing algorithm is required to address IP-based network delay and loss. This work proposes an optimal smoothing method to obtain the best voice quality by Lagrangian lost function which is a trade off between the negative effects of the delay and the loss. It can efficiently solve the mismatch between the capture and the playback clocks. Numerical examples have shown that our proposed method can control the playout time to balance the target delay and loss.

Reference

1. Brazauskas V., Serfling R.: Robust and efficient estimation of the tail index of a one-parameter pareto distribution. North American Actuarial Journal available at http://www.utdallas.edu/~serfling. (2000)
2. Tien P. L., Yuang M. C.: Intelligent voice smoother for silence-suppressed voice over internet. IEEE JSAC, Vol. 17, No. 1. (1999) 29-41
3. Ramjee R., Kurise J., Towsley D., Schulzrinne H.: Adaptive playout mechanisms for packetized audio applications in wide-area networks. Proc. IEEE INFOCOM. (1994) 680-686
4. Jeske D. R., Matragi W., Samadi B.: Adaptive play-out algorithms for voice packets. Proc. IEEE Conf. on Commun., Vol. 3. (2001) 775-779
5. Pinto J., Christensen K. J.: An algorithm for playout of packet voice based on adaptive adjustment of talkspurt silence periods. Proc. IEEE Conf. on Local Computer Networks. (1999) 224-231
6. Liang Y. J., Farber N., Girod B.,: Adaptive playout scheduling using time-scale modification in packet voice communications. Proc. IEEE Conf. on Acoustics, Speech, and Signal Processing, Vol. 3. (2001) 1445-1448
7. Kansal A., Karandikar A.: Adaptive delay estimation for low jitter audio over Internet. IEEE GLOBECOM, Vol. 4. (2001) 2591-2595
8. Anandakumar A. K., McCree A., Paksoy E.: An adaptive voice playout method for VOP applications. IEEE GLOBECOM, Vol. 3. (2001) 1637-1640
9. DeLeon P., Sreenan C. J.: An Adaptive predictor for media playout buffering. Proc. IEEE Conf. on Acoustics, Speech, and Signal Processing, Vol. 6. (1999) 3097-3100
10. Bolot J. C.: Characterizing end-to-end packet delay and loss in the internet. Journal of High-Speed Networks, Vol. 2. (1993) 305-323
11. Huebner F., Liu D., Fernandez J. M.: Queueing Performance Comparsion of Traffic Models for Internet Traffic. GLOBECOM 98, Vol. 1. (1998) 471–476
12. Nobuhiko K. and Kenzo I.: Pure delay effects on speech quality in telecommunications. IEEE JSAC, Vol. 9, No. 4. (1991)
13. Duysburgh B., Vanhastel S., De Vreese B., Petrisor C., and Demeester P.: On the influence of best-effort network conditions on the perceived speech quality of VoIP connections. Proc. Computer Communications and Networks. (2001) 334-339
14. Yamamoto L., Beerends J., KPN Research.L: Impact of network performance parameters on the end-to-end perceived quality. EXPERT ATM Traffic Symposium available at http://www.run.montefiore.ulg.ac.be/~yamamoto/ publications.html. (1997)
15. Melsa P. J. W., Younce R. C., and Rohrs C. E.: Joint impulse response shortening for discrete multitone transceivers. IEEE Trans..Communications, Vol. 44, No. 12. (1996) 1662-1672
16. Hosny N. M., El-Ramly S. H., El-Said M. H.: Novel techniques for speech compression using wavelet transform. The International Conference on Microelectronics. (1999) 225-229