Application-aware Retransmission Design for VoIP Services in BWA Networks

Sung-Min Oh * and Jae-Hyun Kim**

* ETRI(Electronics and Telecommunications Research Institute), Korea

** School of Electrical and Computer Engineering, Ajou University

smoh@etri.re.kr, jkim@ajou.ac.kr

***Abstract*— To improve the user-perceived quality of service (QoS) performance of voice over Internet protocol (VoIP) services, a retransmission of a packet experienced an error is needed in broadband wireless access networks. This means that the number of wireless resources to successfully transmit a packet can be increased by the retransmitted packet. For this reason, there can be a trade-off between the QoS performance improvement and the wireless resource efficiency enhancement with respect to the maximum number of retransmissions. This paper has formulated the cross-layer optimization problem to maximize the wireless resource efficiency with a constraint of the user-perceived QoS performance. This paper has driven the R-value for the user-perceived QoS performance because it includes most of impairments that can be generated during transmitting a voice frame from mouth to ears. Due to this feature, we can directly assess the QoS performance of VoIP services from a viewpoint of users. In addition, we can clearly design the optimal maximum number of retransmissions that can provide the satisfied quality of a VoIP service. By the numerical results, the optimal maximum number of retransmissions to provide the fair quality of a VoIP service under given conditions for G.711, G.729 and G.723.1 are 1, 3 and 4, respectively.**

***Keywords*— Cross-layer optimization, user-perceived QoS performance, truncated ARQ, and VoIP services.**

I. INTRODUCTION

Voice over Internet protocol (VoIP) services have been considered as one of the most important services in the next generation wireless systems. To provide VoIP services in wireless networks, there are two main research issues such as the quality of service (QoS) provisioning and the efficient use of wireless resources. For this reason, several papers relevant to the research issues have been published [1] - [5]. In these papers, the authors including us have studied on VoIP scheduling algorithms that can efficiently use the wireless resources without the increase of an access delay in wireless networks. However, these papers have not considered a packet loss rate (PLR) with respect to a QoS provisioning of VoIP services.

Actually, if we consider a QoS performance of VoIP services from the viewpoint of a listener, an end-to-end PLR can be more important factor than an end-to-end delay. In [6] and [7], the authors have evaluated the user-perceived QoS performance using the R-value. The R-value has been widely used to assess the user-perceived QoS performance because it includes most of impairments from mouth to ears [6] - [8]. The authors have mentioned that the R-value is more sensitive to an end-to-end PLR than an end-to-end delay. This implies that a PLR can be an important performance metric for guaranteeing a quality of VoIP services in wireless networks. In [9], we have simulated the user-perceived QoS performance while a user moves from a center of a cell to a boundary of the cell. The simulation results have shown that nearly all users dissatisfy the quality of VoIP services due to the increase of a PLR at the boundary of the cell, although the average access delay is maintained under ten milliseconds.

In [9], we have also verified that truncated automatic repeat request (ARQ) can help to improve the user-perceived QoS performance at the low signal to noise ratio (SNR) even though truncated ARQ can increase an average access delay. The reason of this is that truncated ARQ can maintain the bound of an access delay with decreasing a PLR by the maximum number of retransmissions ($N_{\max}$). Therefore, the user-perceived QoS performance can be improved as we increase $N_{\max}$ by a certain point. After the certain point, the user-perceived QoS performance can be deteriorated because the bound of an access delay is increased by more than a threshold that can affect to the QoS performance. For this reason, we can regard that the certain point is the optimal $N_{\max}$. However, this derivation is incomplete, because it has not taken the efficient use of wireless resources into account.

There can be a disadvantage that a retransmission of a VoIP packet can increase the number of wireless resources required to successfully transmit a VoIP packet. This means that the maximum number of supportable VoIP users under a limited capacity of a wireless system, which is called as VoIP capacity in this

paper, can be decreased as N_{max} is increased. Consequently, there is a tradeoff between the QoS performance improvement and the wireless resource efficiency enhancement with respect to N_{max}.

To solve the problem, we model the R-value and VoIP capacity with respect to truncated ARQ and formulate a cross-layer optimization problem with respect to N_{max} in Section III. Section IV presents performance analysis results and the optimal N_{max} according to VoIP speech codecs. In Section V, we conclude this paper.

Contributions of this paper are as follows:

- To design the cross-layer optimization, we have modeled the R-value with respect to truncated ARQ with considering system parameters of the whole layer.
- To the best of our knowledge, we have firstly modeled and solved the cross-layer optimization problem with a constraint of the user-perceived QoS performance.
- The cross-layer optimization format can be easily solved by using mixed integer non linear programming (MINLP). Thus, it can be widely used in various researches for the next generation wireless systems.
- We have considered error-correction capabilities of VoIP speech codecs by using the R-value. With this contribution, we have finally presented the optimal N_{max} according to the kinds of VoIP speech codecs. The optimal N_{max} can be applied to various systems that include truncated ARQ.

II. Related Works

Since truncated ARQ is one of the main wireless channel dependent technologies in the next generation wireless systems, a number of papers relevant to it have been published [10] - [16]. In [10] - [12], the authors have analyzed ARQ and hybrid ARQ (HARQ) while considering various communication environments such as wireless fading channel, adaptive modulation and coding (AMC), code combining and so on. However, the authors have merely focused on the performance analysis of the physical layer and the data link layer. In [13], the authors have studied the cross-layer design for AMC and truncated ARQ by using a two-dimensional Markov model. In [14], the authors have investigated truncated ARQ with multiuser scheduling. However, in these two papers, the authors have not drawn the optimal N_{max}. In [15] and [16], the authors have investigated a cross-layer design for the QoS provisioning and they have presented the optimal N_{max}. However, in these papers, they have regarded that a target PLR and a target delay are independent and constant. Unfortunately, the target PLR is correlated with the target delay with respect to a certain QoS level [7]. This means that if we would like to maintain a certain QoS level, the target PLR and the target delay can be complementally changed. Therefore, the optimal N_{max} presented in [15] and [16] may not be optimal in terms of the wireless resource efficiency and the user-perceived QoS performance.

In order to compensate these weaknesses as mentioned above, we derive the R-value for the cross-layer optimization. The reason of this is that the R-value can be considered the correlation of the PLR and the delay because it is presented by a relation of the end-to-end PLR and the end-to-end delay. Particularly, the R-value has another merit that it can imply error-correction capabilities of VoIP speech codecs. For these reasons, we can clearly assess the QoS performance from a viewpoint of a listener by using the R-value.

In [6] and [7], Sengupta et al. have evaluated several schemes to improve a QoS performance of VoIP services using the R-value. However, they have not considered the VoIP capacity and hence the optimal N_{max} has not been driven in these papers. Consequently, in this paper, we design the cross-layer optimization problem to maximize the VoIP capacity with a constraint of the R-value. Finally, we find out the optimal N_{max} according to VoIP speech codecs.

III. Modeling

In this paper, we consider the end-to-end reference network architecture as shown in Fig. 1. To evaluate the user-perceived QoS performance with respect to truncated ARQ, we assume that there is no congestion in the backbone network and the wired link capacity is sufficiently large to send a packet. By these assumptions, we can define that the time to send a packet through the backbone network is $d_{backbone}$ (constant) and that there is no packet loss in the backbone network. This means that the PLR in end-to-end networks ($P_{d,net}$) can be equal to the packet drop rate in the access network.

In addition, since VoIP speech codecs need the time to encode or decode a voice frame, we assume that VoIP packets are delayed by $d_{encoding}$ and $d_{decoding}$ (constant). This assumption is reasonable because voice frames are periodically generated [1]. At the receiver side, VoIP speech codecs generally have a playout buffer to compensate a network jitter or delay. For this reason, a packet can be delayed by $d_{playout}$ in the playout buffer. $d_{playout}$ is assumed as a constant [6]. A packet can be dropped if the delay elapsed to send a packet through the end-to-end network is longer than the playout buffer size (δ) in seconds. In this paper, the packet drop rate in the playout buffer is defined as $P_{d,buf}$.

For the access network, we consider a system model that is made of functions described in Fig. 2. The system model consists of the higher layer, the data link layer, and the physical layer.

At the higher layer, this paper considers the operation procedure for a VoIP service from the application layer

to the Internet protocol (IP) layer. In general, a VoIP service is supported by the real-time protocol (RTP), user-datagram protocol (UDP), and IP. In addition, the size of a VoIP packet is much smaller than that of the maximum transmission unit (MTU). For this reason, a VoIP packet can be delivered from the application layer to the IP layer without fragmentation. In the IEEE 802.16 standards, the payload header suppression (PHS) was defined to enhance the system efficiency [17]. We also apply the PHS to the system in this paper; thus the input module transmits a service data unit (SDU), which consists of a suppressed header (SH) and a VoIP packet, to the data link layer. The SH includes the suppressed information of the RTP header, UDP header, and IP header [1], [4].

At the data link layer, the selective ARQ is implemented. The processing unit is a protocol data unit (PDU) that encapsulates the SDU received from the higher layer. The PDU includes header, payload, and cyclic redundancy check (CRC) to facilitate error detection. If an error is detected in the PDU, a retransmission request is generated by the ARQ generator. It is sent to the ARQ controller at the transmitter through the feedback channel [15]. The ARQ controller arranges the retransmitting packet that is stored in the buffer. In this paper, we assume that the feedback messages are error-free.

With the ARQ modules, the scheduler module is also applied at the data link layer. The scheduler module includes a call admission control (CAC) function. The CAC can determine whether a VoIP service flow can be supported or not while considering the system capacity. This means that there is no queuing delay. In addition, we apply a semi-persistent resource allocation algorithm proposed in [4] and [17] to the scheduler module. Since the semi-persistent resource allocation algorithm can reserve wireless resource for VoIP traffic, there is no access delay to request an uplink resource. Here, the access delay is the time elapsed to transmit a packet from a transmitter to a receiver.

To retransmit a packet which experiences an error, the transmitter also needs to obtain wireless resources for the retransmitting packet. As shown in Fig. 3, the retransmitting packet has to be delayed by T_d seconds that is a resource allocation interval (RAI), because the transmitter can request the additionally required resource by polling or piggybacking for every RAI [4], [17].

At the physical layer, we assume that the average PDU error rate (PER) is a variable (p) because the PER is not a control parameter. It is dependent on the wireless channel condition and the wireless channel dependent technologies such as a forward error correction, modulation, interleaving, and so on.

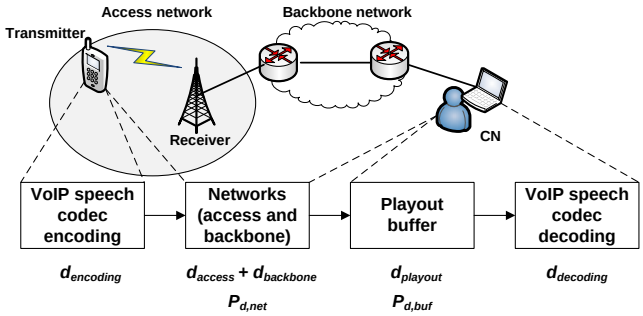

Figure 1. End-to-end reference network architecture

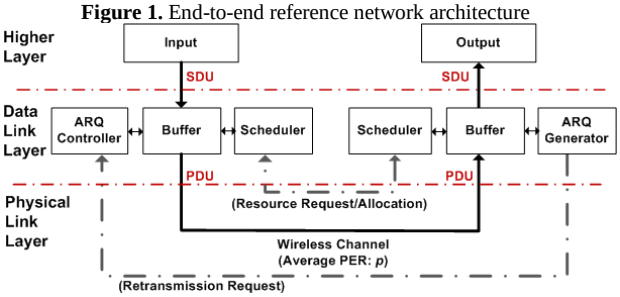

Figure 2. System model in wireless (access) networks

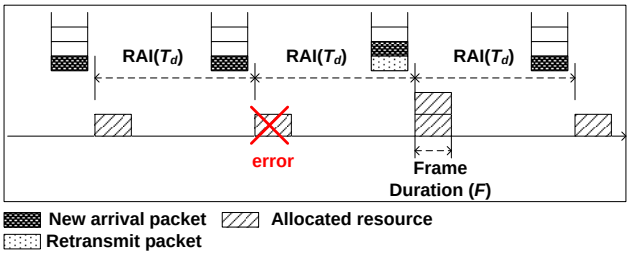

Figure 3. Resource allocation operation for a retransmitting packet

IV. Performance Analysis

The goal of this paper is to obtain the optimal $N_{\max}$ in terms of the user-perceived QoS performance and the wireless resource efficiency. In order to achieve the performance analysis, three steps are performed in this paper. Firstly, we present the performance analysis model of the data link layer. Secondly, we model the user-perceived QoS performance with respect to $N_{\max}$ and p. Finally, we design the cross-layer optimization problem and obtain the optimal $N_{\max}$.

A. Performance analysis in the data link layer

A PDU is retransmitted by the ARQ module when the PDU has an error. To consider this feature, let us denote that the number of PDU transmissions to successfully send a PDU is x. Thus, the probability mass function (*pmf*) in terms of x can be defined as

$$P[X = x] = \frac{(1-p)\cdot p^{x-1}}{\sum_{i=1}^{N_{\max}+1}(1-p)\cdot p^{i-1}} = \frac{(1-p)\cdot p^{x-1}}{1-p^{N_{\max}+1}}, \quad (1)$$

where x = 1, 2, …, $N_{\max}$+1. For this reason, the average number of PDU transmissions is

$$\overline{X}(p, N_{\max}) = E[X] = \sum_{x=1}^{N_{\max}+1} x \cdot P[X = x] = \frac{1}{1-p} - \frac{(N_{\max}+1) \cdot p^{N_{\max}+1}}{1-p^{N_{\max}+1}} . \quad (2)$$

In this paper, there is no queuing delay because of the CAC function and the elapsed time to retransmit a PDU is T_d in seconds. Thus, the average access delay ($\overline{D}_a$) can be defined as

$$\overline{D}_a(p, N_{\max}) = F + \left(\overline{X}(p, N_{\max}) - 1\right) \times T_d , \quad (3)$$

where F is the size of a MAC frame in seconds.

If a PDU continuously experiences an error until the maximum number of transmission ($N_{\max}$+1) has been reached, the PDU is dropped. By this feature, the PDU drop rate ($P_{d,net}$) can be given by

$$P_{d,net}(p, N_{\max}) = p^{N_{\max}+1} . \quad (4)$$

B. Performance analysis for the playout buffer

Since VoIP services are real-time inter-communication services, a long-delayed packet can be unuseful. For this reason, there is a playout buffer in the application layer and the long-delayed packet is dropped in the playout buffer. In this subsection, we model the packet drop probability at the playout buffer.

For an instantaneous x, the access delay of the data link layer can be given by

$$D_a(x) = F + (x-1) \times T_d . \quad (5)$$

As shown in Fig. 1, the end-to-end delay can be defined as

$$d(x) = d_{encoding} + D_a(x) + d_{backbone} + d_{playout} + d_{decoding} . \quad (6)$$

At the playout buffer, a packet can be dropped if the delay of the packet is longer than the playout buffer size (δ). A condition for this feature is as follows:

$$d_{encoding} + D_a(x) + d_{backbone} > \delta . \quad (7)$$

By using (5) and (7), a condition for the number of PDU transmissions to be dropped at the playout buffer can be given as

$$x > \frac{\delta - d_{encoding} - d_{backbone} - F}{T_d} + 1 . \quad (8)$$

Here, by using (8), we can define the maximum number of PDU transmissions in the access network as (9) to avoid being dropped at the playout buffer.

$$\delta_{th} = \left\lfloor \frac{\delta - d_{encoding} - d_{backbone} - F}{T_d} \right\rfloor + 1 , \quad (9)$$

where $\lfloor a \rfloor$ means the largest integer value that is smaller than a. In order to be dropped at the playout buffer, the packet has to be not lost in the access network and the number of PDU transmissions in the access network is also larger than δ_{th}. By this feature, we can model the packet drop rate at the playout buffer as follows:

$$P_{d,buf}(p, N_{\max}) = \frac{(1-p) \cdot P\left[X > \delta_{th}\right]}{1-p^{N_{\max}+1}} = \begin{cases} \dfrac{p^{\delta_{th}} - p^{N_{\max}+1}}{1-p^{N_{\max}+1}} & , \delta_{th} \le N_{\max}+1 \\ 0 & , \delta_{th} > N_{\max}+1 \end{cases} . \quad (10)$$

C. Performance analysis for the R-value

To analyze the user's satisfaction of VoIP services, we consider the R-value [6]. The R-value indicates a quality of a VoIP service calculated by using features of a VoIP speech codec and performance metrics of the end-to-end network such as the end-to-end delay and the end-to-end PLR. The derivation process of the R-value is as follows:

The average end-to-end delay means the average delay of a packet which is delivered to ears of a listener. By this feature and (6), the average end-to-end delay can be defined as

$$\begin{aligned} \overline{d}(p, N_{\max}) &= E[d(x)] = (1 - P_{d,buf}(p, N_{\max})) \cdot \sum_{x=1}^{N_{\max}+1} d(x) \cdot P[X = x] \\ &= (1 - P_{d,buf}(p, N_{\max})) \cdot \left\{ C \cdot \sum_{x=1}^{N_{\max}+1} P[X = x] + T_d \cdot \sum_{x=1}^{N_{\max}+1} x \cdot P[X = x] \right\} \\ &= (1 - P_{d,buf}(p, N_{\max})) \cdot \left\{ C + \overline{X}(p, N_{\max}) \cdot T_d \right\} \end{aligned} . \quad (11)$$

where $C = d_{encoding} + d_{backbone} + d_{playout} + d_{decoding} + F - T_d$.

Since the packet loss through the end-to-end networks indicates that a packet is dropped in the access network or the playout buffer, we can define the end-to-end PLR as

$$e(p, N_{\max}) = P_{d,net}(p, N_{\max}) + (1 - P_{d,net}(p, N_{\max})) \cdot P_{d,buf}(p, N_{\max}) . \quad (12)$$

In [6] and [8], the R-value with default values is given by

$$R = 94.2 - I_e - I_d , \quad (13)$$

where I_e is an equipment impairment factor associated with the losses due to the codec and network and I_d presents the impairment caused by the delay. In (13), I_e is given by

$$I_e(p, N_{\max}) = G_1 + G_2 \cdot \ln(1 + G_3 \cdot e(p, N_{\max})) , \quad (14)$$

where G_1 is a constant related to encoding, and G_2 and G_3 mean the impact of loss for a given codec [8]. In (13), I_d is defined as

$$I_d(p, N_{\max}) = 0.024 \cdot \bar{d}(p, N_{\max}) + 0.11 \cdot (\bar{d}(p, N_{\max}) - 177.3), \times H(\bar{d}(p, N_{\max}) - 177.3) \quad (15)$$

where $H(x)$ is an indicator function [8]. $H(x) = 0$ if $x < 0$ otherwise, $H(x) = 1$. Thus, the R-value can be presented a function of p and N_{max} by (13), (14), and (15).

D. Performance analysis for VoIP capacity

A traffic rate and a packet-generation-interval (PGI) of VoIP services can be variable according to the voice activity such as a talk-spurt and a silent-period. Considering this property, we obtain the VoIP capacity in this subsection. In this paper, we define the unit of the allocated resource as a slot.

The VoIP capacity is the maximum number of supportable users under given slots in a MAC frame. To obtain the VoIP capacity, we define the number of slots allocated for a packet during a talk-spurt and a silent-period as follows:

$$S_{on} = \left\lceil \frac{L_{PH} + L_{SH} + L_{TPK} + L_{CRC}}{B} \right\rceil, \quad (16)$$

$$S_{off} = \left\lceil \frac{L_{PH} + L_{SH} + L_{SPK} + L_{CRC}}{B} \right\rceil, \quad (17)$$

where L_{PH}, L_{SH}, L_{TPK}, L_{SPK}, and L_{CRC} are the number of bytes of a PDU header, a SH, a packet in a talk-spurt, a packet in a silent-period, and a CRC, respectively. In addition, B is the number of bytes of a slot. Using (16) and (17), the average number of slots allocated for a packet for every a PGI in a talk-spurt can be given by

$$\bar{S} = S_{on} \cdot T_{on} + S_{off} \cdot \frac{T_{TPGI}}{T_{SPGI}} \cdot T_{off}, \quad (18)$$

where T_{on} and T_{off} indicate the duration of the talk-spurt and the silent-period, respectively. In addition, the summation of T_{on} and T_{off} is one second. T_{TPGI} and T_{SPGI} are the PGI of a talk-spurt and a silent-period, respectively. By using (2) and (18), we can obtain the VoIP capacity as

$$\bar{m}(p, N_{\max}) = \frac{T_{TPGI}}{F} \times \frac{S_{tot}}{\bar{S} \cdot \bar{X}(p, N_{\max})}, \quad (19)$$

where S_{tot} is the total number of slots in a MAC frame. In (19), the left term on the right side means the number of MAC frames during a PGI of a talk-spurt, and the right term indicates the average number of supportable users for every the PGI of a talk-spurt.

E. Cross-layer optimization problem

By (13) and (19), the R-value and the VoIP capacity are given by functions of p and $N_{\max}$. However, since the PER (p) is not a control parameter, the averaging process of the R-value and the VoIP capacity in terms of p is needed as described in Appendix A.

The goal of this paper is to find out the optimal $N_{\max}$ to maximize the VoIP capacity with a constraint of the R-value. For this reason, we can formulate a cross-layer optimization problem as follows:

$$\max_{N_{\max}} \quad \bar{m}(N_{\max}) \quad (20)$$
$$s.t.$$
$$\bar{R}(N_{\max}) \geq R_{th}, \quad N_{\max} \in \{0, 1, 2, ..., Q\},$$

where R_{th} indicates the threshold of the R-value and Q is a maximum value of $N_{\max}$. Here, the objective function is a convex, the constraint function is a concave, and $N_{\max}$ is an integer. For this reason, we can solve the optimization problem by using the MINLP [18].

V. Numerical Results

In this section, we analyze the performance in the data link layer and the end-to-end networks. In addition, we present the numerical results of the R-value and the VoIP capacity in order to analyze the user-perceived QoS performance and the wireless resource utilization. Parameters for performance analysis are listed in Table I. In this paper, we consider the uplink transmission of an orthogonal frequency division multiple access (OFDMA)/time division duplex (TDD) system with 10 MHz and 1024 fast Fourier transform (FFT). In [17], there are 840 subcarriers and 18 symbols in an uplink subframe, and a slot consists of 24 subcarriers and 3 symbols. For this reason, S_{tot} can be calculated as 210 slots. In [19], the authors have measured a playout buffer size. In the paper, they have presented that a playout buffer size is about 300 msec when application programs such as Skype and Google Talk are applied. For this reason, we define δ as 300 msec in this paper. Since T_d depends on the kinds of VoIP speech codecs, we define that T_d is equal to T_{TPGI}. We refer the other parameters at related papers and standards.

TABLE 1. Parameters for Performance Analysis

Parameters	Value
Total number of resources (S_{tot})	210 slots [17]
MCS level	QPSK 1/2
T_{on}, T_{off}	0.4, 0.6 [1]
L_{PH}, L_{SH}, L_{CRC}	6 bytes, 3 bytes, 4 bytes [17]
$d_{encoding}$, $d_{backbone}$, $d_{decoding}$, $d_{playout}$	20 msec [6], 20 msec [20], 20 msec [6], 60 msec [6]
Playout buffer size (δ)	300 msec [19]
Delay to retransmit a packet (T_d)	T_{TPGI}

MAC frame size (F)		5 msec [17]
G.723.1 [22]	G_1, G_2, G_3	19, 37.4, 6 [21]
	T_{TPGI}, T_{SPGI}	30 msec, 160 msec
	L_{TPK}, L_{SPK}	20 bytes, 2 bytes
G.729 [23]	G_1, G_2, G_3	11, 40, 10 [6]
	T_{TPGI}, T_{SPGI}	10 msec, 160 msec
	L_{TPK}, L_{SPK}	10 bytes, 2 bytes
G.711 [24]	G_1, G_2, G_3	0, 30, 15 [6]
	T_{TPGI}, T_{SPGI}	20 msec, 160 msec
	L_{TPK}, L_{SPK}	160 bytes, 2 bytes

A. Performance analysis in the access network

Fig. 4 represents the PDU drop rate for G.711 in the access network with respect to p and N_{max}. The case, that N_{max} is equal to zero, means that the ARQ is not applied to the system. As shown in Fig. 4, the PDU drop rate is linearly increased by 0.5 as p increases when the ARQ is not applied to the system. The reason is that all packets which experience an error are dropped when the ARQ is not applied to the system. However, the PDU drop rate can be dramatically decreased as N_{max} increases. Specifically, the PDU drop rate is about zero when N_{max} is larger than 4. Unfortunately, the access delay can be increased as N_{max} increases. Since VoIP services are sensitive to the delay, we need to analyze the average access delay.

Fig. 5 indicates the average access delay for G.711 with respect to p and N_{max}. The average access delay is logarithmically increased by about 25 msec as N_{max} increases up to 7 when p is about 0.5. The increased delay is due to PDU retransmissions. However, the increasing rate is decreased while N_{max} increases, because the number of PDUs, which are successfully transmitted before reaching N_{max}, increases.

From Figs. 4 and 5, we can grasp that there is a tradeoff between the average access delay and the PDU drop rate with respect to N_{max}. However, we cannot exactly assess the user-perceived QoS performance while considering only the performance in the data link layer. The reason of this is as follows. The effect of the user-perceived PLR can be different according to the kinds of VoIP speech codecs although the PDU drop rate in the data link layer is the same, because each VoIP speech codec has a specific error recovery function. In addition, there is a playout buffer in the application layer. For this reason, a long-delayed packet in the data link layer may be dropped in the playout buffer. This means that the PLR can be increased by the long-delayed packets from a viewpoint of a receiver. Therefore, we need to analyze the user-perceived QoS performance.

B. End-to-end performance analysis

Figs. 6 and 7 describe the average end-to-end delay and the packet drop rate for G.711 in the playout buffer, respectively. The symbols indicate the simulation results and the dashed lines mean the analytical results. As shown in Figs. 6 and 7, we can confirm that the simulation results are similar to the analytical results.

In Fig. 6, the average end-to-end delay is linearly increased by 136 and 142 msec as p increases up to 0.5 when N_{max} is 2 and 4, respectively. In case of N_{max} = 15, the average end-to-end delay is almost the same as that of N_{max} = 4 when p is smaller than 0.3. The reason of this is that there is almost no packet retransmitted more than four times in the data link layer. We can analyze this situation from Fig. 4. When p is larger than 0.3, the average end-to-end delay of N_{max} = 15 is longer than that of N_{max} = 4. From this result, we can estimate that there may be a packet dropped at the playout buffer.

As shown in Fig. 7, a few packets can be dropped at the playout buffer when N_{max} is 15 and p is larger than 0.3. Actually, in the case of G.711, δ_{th} is 13 under given conditions as listed in Table 1. For this reason, a packet can be dropped when N_{max} is larger than 14. However, we are generally interested in N_{max} which is smaller than 14. Therefore, we can assume that there is no dropped packet at the playout buffer in this paper. This means that the end-to-end packet drop rate depends on the PDU drop rate.

C. R-value and VoIP capacity

While considering the end-to-end performance, we analyze the R-value and the VoIP capacity. The R-value can be mapped with the user satisfaction of a VoIP service as follows:

- R-value (90) - excellent service quality
- R-value (80) - good service quality
- R-value (70) - fair service quality
- R-value (60) - poor service quality
- R-value (50) - bad service quality

As shown in Fig. 8, the R-value of G.711 is seriously affected by p when the ARQ is not applied to the system. Especially, we can analyze that it is difficult to provide the fair quality of a VoIP service of G.711 when p is larger than 0.1. As mentioned in Appendix A, p is about 0.5 at the boundary of a cell. Thus, we can infer that it is difficult to guarantee a QoS performance without truncated ARQ. The R-value is logarithmically increased as N_{max} increases, and it is larger than 70 when N_{max} is 3 although p increases up to 0.5. This means that the fair quality of a VoIP service can be provided if we select N_{max} as 3. However, the increase of N_{max} can cause the degradation of the wireless resource efficiency. For this reason, the N_{max} may not be optimal. Therefore, we consider the VoIP capacity to assess the wireless resource efficiency.

Fig. 9 presents the VoIP capacity of G.711. As shown in Fig. 9, the VoIP capacity is logarithmically decreased as N_{max} increases. Especially, the VoIP

capacity is decreased by 43 % as $N_{\max}$ increases up to 7 when p is 0.5. The decrease of the VoIP capacity is due to the increase of the number of PDU transmissions to successfully send. This indicates that p is another factor to affect to the VoIP capacity. For this reason, we cannot directly drive the optimal $N_{\max}$ from Figs. 8 and 9. Since p is not a control parameter, we perform an averaging process of the R-value and the VoIP capacity in terms of p as described in Appendix A.

D. Optimal $N_{\max}$

Table II represents the optimal $N_{\max}$ of the cross-layer optimization and QoS requirement based according to VoIP speech codecs. The cross-layer optimization is the proposed scheme to find out the optimal $N_{\max}$ in this paper. The QoS requirement based is the conventional scheme used in [15] and [16]. The conventional scheme exploits the predefined QoS requirement to choose the optimal $N_{\max}$. In Table II, the $N_{\max}$ of this scheme is selected by using the delay requirement. In this paper, we define the delay requirement as 50 msec [25]. Thus, $N_{\max}$ of the QoS requirement based can be calculated by using the delay requirement and T_d. In addition, we consider the R-value (70) as the minimum value of the satisfied QoS performance in this paper.

In the case of G.711, the average R-value and the average VoIP capacity of the QoS requirement based are fixed as 55.30 and 77.00. On the other hand, the cross-layer optimization can dynamically select the optimal $N_{\max}$ for various goals in the system design. The goal of a system can be adapted by selecting R_{th}. If we would like to design the system to provide better QoS performance, we can select R_{th} as 90. Then, the optimal $N_{\max}$ is 5 by the cross-layer optimization scheme, and the average VoIP capacity and the average R-value are 51.75 and 90.02, respectively. If we would like to design the system to support more users with the satisfied QoS performance, we can select R_{th} as 70. Then, the optimal $N_{\max}$ is 1, and the average VoIP capacity and the average R-value are 59.59 and 70.46, respectively.

The case of G.729 is similar to that of G.711. However, there is no optimal $N_{\max}$ when R_{th} is 90. This is caused by the feature of the VoIP speech codec. Since G.729 delivers voice frames with a low data rate in order to enhance the network resource efficiency, it relatively has a low QoS performance. For this reason, G.729 has a limitation of the QoS performance.

In the case of G.723.1, the QoS performance for the QoS requirement based is seriously degraded. The reason of this is that the QoS requirement based does not consider the relation of the delay and the PLR for the QoS performance. In the QoS requirement based, since the retransmission delay of G.723.1 is quite long for the fixed delay requirement, it selects $N_{\max}$ as 1. However, the decrease of the PLR can compensate the QoS performance degradation by the increased delay. By considering this property, the cross-layer optimization can select the optimal $N_{\max}$ as 4. Under this condition, the average R-value can be increased to 70.53. Although the average VoIP capacity of the cross-layer optimization is decreased by 24 % compared to that of the QoS requirement based, we can infer that the cross-layer optimization is superior to the QoS requirement based because the QoS provisioning is prior to the wireless resource efficiency in the system design.

VI. CONCLUSION

In this paper, we have analyzed the user-perceived QoS performance and the VoIP capacity with respect to the maximum number of retransmissions for VoIP services. Numerical results have shown that the ARQ is needed to improve the user-perceived QoS performance of VoIP services. Especially, we have found out the optimal $N_{\max}$ by maximizing the VoIP capacity while providing the satisfied quality of a VoIP service.

In this paper, we have shown that the system performance and the QoS performance can be improved by considering specific features of VoIP services in the application layer. This approach is valuable because it can be applied to various wireless systems.

ACKNOWLEDGMENT

This work was supported by the IT R&D program of MKE/KEIT. [10039176, Researches on the Gbps wireless technology for providing multimedia services to the group of subscribers in a platform moving faster than 300km/h]

APPENDIX A

THE AVERAGING PROCESS OF THE R-VALUE AND THE VOIP CAPACITY IN TERMS OF *P*

For the averaging process in terms of p, we assume two conditions as follows:

Under given parameters as listed in Table I, δ_{th} is equal to 8. This means that $P_{d,buf}$ is zero when $N_{\max}$ is smaller than 8. In general, we are interested in a value of $N_{\max}$ which is smaller than 8. Therefore, we assume that $P_{d,buf}$ is zero. In this paper, we call this assumption as A1).

In [17], the minimum signal noise ratio (SNR) at a cell edge is about 5 dB ~ 8 dB by a table of the modulation and coding scheme (MCS) level with a handover. In addition, in order to obtain an average PER according to a SNR, we have simulated a wireless system that includes several technologies, such as randomization, forward error correction (FEC), interleaving, and modulation. By the simulation results, we have found that the average PER is smaller than about 0.5 when the SNR is higher than 8 dB. Therefore,

we can assume that a range of an average PER is from 0 to 0.5. By this assumption, we can define the range of p as $[0, p_t]$ where p_t is the highest value of p in the system and $p_t < 1$.

The average number of PDU transmissions can be given by

$$\overline{X}(N_{\max}) = \frac{1}{p_t}\int_0^{p_t} \frac{1}{1-p}\,dp - \frac{1}{p_t}\int_0^{p_t} \frac{(N_{\max}+1)\cdot p^{N_{\max}+1}}{1-p^{N_{\max}+1}}\,dp \quad . \tag{21}$$

In (21), solving the integral of the second term on the right side is complex. Thus, we approximately solve the integral by using Taylor's expansions. Therefore, (21) can be driven as

$$\overline{X}(N_{\max}) \approx \frac{1}{p_t}\cdot\left\{\ln\left(\frac{1}{1-p_t}\right) - \left(\frac{N_{\max}+1}{N_{\max}+2}\cdot p_t^{N_{\max}+2} + \frac{N_{\max}+1}{2N_{\max}+3}\cdot p_t^{2N_{\max}+3}\right)\right\} \quad . \tag{22}$$

In (11), we can obtain $\overline{d}(N_{\max})$ by applying A1) and substituting $\overline{X}(N_{\max})$ for $\overline{X}(p, N_{\max})$. Thus, we can drive $\overline{I_d}(N_{\max})$ by substituting $\overline{d}(N_{\max})$ for $\overline{d}(p, N_{\max})$ in (15). By using 4) and A1), (14) for small values of p can be written as

$$\overline{I_e}(N_{\max}) = \frac{1}{p_t}\int_0^{p_t} G_1\,dp + \frac{1}{p_t}\int_0^{p_t} G_2\cdot\ln\left(1+G_3\cdot p^{N_{\max}+1}\right)dp \quad . \tag{23}$$

In (23), solving the second term on the right side is very complex. Thus, we solve the integral by using the approximation method. Since $p^{N\max+1}$ is lower than 0.125 when $N_{\max} > 1$, we can apply Taylor's expansions. Therefore, (23) can be driven as

$$I_e(0) = G_1 + \frac{G_2}{p_t}\cdot\left(\frac{\ln(1+G_3\cdot p_t)\cdot(1+G_3\cdot p_t)}{G_3} - p_t\right), \tag{24}$$

$$I_e(1) = G_1 + \frac{G_2}{p_t} \times\left(p_t\cdot\ln(1+G_3\cdot p_t^2) - 2\cdot p_t + \frac{2}{\sqrt{G_3}}\cdot\arctan\left(p_t\cdot\sqrt{G_3}\right)\right), \tag{25}$$

$$I(N_{\max}) \approx G_1 + G_2\cdot G_3\cdot\frac{p_t^{N_{\max}+1}}{N_{\max}+2}, \quad N_{\max} \geq 2\,. \tag{26}$$

Consequently, we can obtain the average of the R-value in terms of p by using (13), (24), (25), and (26). In addition, in (19), we can drive the average of the VoIP capacity in terms of p by substituting $\overline{X}(N_{\max})$ in (22) for $\overline{X}(p, N_{\max})$.

REFERENCES

[1] H. Lee, T. Kwon, and D.-H. Cho, "An Enhanced Uplink Scheduling Algorithm Based on Voice Activity for VoIP Services in IEEE 802.16d/e System," *IEEE Commun. Lett.*, vol. 9, pp. 691-693, Aug. 2005.

[2] H. Lee, T. Kwon, and D.-H. Cho, "An Efficient Uplink Scheduling Algorithm for VoIP Services in IEEE 802.16 BWA Systems," in *Proc. IEEE Vehicular Technology Conf*., vol. 5, pp. 3070 – 3074, 2004.

[3] H. Lee, H.-D. Kim, and D.-H. Cho, "Smart Resource Allocation Algorithm Considering Voice Activity for VoIP Services in Mobile-WiMAX System," *IEEE Trans. on Wireless Commun.*, vol. 8, pp. 4688 – 4697, 2009.

[4] S.-M. Oh, S. Cho, J.-H. Kwun, and J.-H. Kim, "VoIP Scheduling Algorithm for AMR Speech Codec in IEEE 802.16e/m System," *IEEE Commun. Lett.*, vol. 12, no. 5, pp. 374 – 376, May 2008.

[5] S.-M. Oh, S. Cho, J.-H. Kim, and J. Kwun, "An Efficient Uplink Scheduling Algorithm with Variable Grant-Interval for VoIP Service in BWA Systems," *IEICE Trans. Commun.*, vol. E91-B, no. 10, Oct. 2008.

[6] S. Sengupta, M. Chatterjee, and S. Ganguly, "Improving Quality of VoIP Streams over WiMAX," *IEEE Trans. on Computers*, vol. 57, no. 2, Feb. 2008.

[7] S. Sengupta et al., "Improving R-Score of VoIP Streams over WiMAX," in *Proc. IEEE Int'l Conf. Commun.* (*ICC '06*), vol. 2, 2006, pp. 866 – 871.

[8] ITU-T G.107, "The E-model, A Computational Model for Use in Transmission Planning," 2000.

[9] S.-M. Oh and J.-H. Kim, "User-Perceived QoS Performance Enhancement for VoIP Services in IEEE 802.16 Systems," in *Proc. Pervasive Wireless Network* (*PWN '10*), Mannheim, Germany, 2010.

[10] E. Malkamaki and H. Leib, "Performance of Truncated Type-II Hybrid ARQ Schemes with Noisy Feedback over Block Fading Channels," *IEEE Trans. on Commun.*, vol. 48, no. 9, pp. 1477 – 1487, Sep. 2000.

[11] H.-C. Yang and S. Sasankan, "Joint AMC/ARQ Transmission in Wireless TDMA Systems and Its Performance Analysis," in *Proc. Wireless Comm. and Networing Conf.* (*WCNC '06*), vol. 3, 2006, pp. 1299 – 1304.

[12] Q. Chen and P. Fan, "Performance Analysis of Hybrid ARQ with Code Combining over Interleaved Rayleigh Fading Channel," *IEEE Trans. on Vehicular Technology*, vol. 54, no. 3, pp. 1207 – 1214, 2005.

[13] J. Ramis, L. Carrasco, and G. Femenias, "A Two-dimensional Markov Model for Cross-layer Design in AMC/ARQ-based Wireless Networks," in *Proc. IEEE GLOBECOM 2008*, 2008, pp. 1 – 6.

[14] X. Wang et al., "Incorporating Retransmission in Quality-of-Service Guaranteed Multiuser Scheduling Over Wireless Links," *IEEE Trans. on Vehicular Technology*, vol. 58, no. 8, pp. 4388 – 4397, Oct. 2009.

[15] Q. Liu, S. Zhou, and G. B. Giannakis, "Cross-Layer Combining of Adaptive Modulation and Coding With Truncated ARQ Over Wireless Networks," *IEEE Trans. on Wireless Commun.*, vol. 3, no. 5, pp. 1746 – 1755, Sep. 2004.

[16] X. Wang et al., "Analyzing and Optimizing Adaptive Modulation and Coding Jointly With ARQ for QoS-Guaranteed Traffic," *IEEE Trans. on Vehicular Technology*, vol. 56, no. 2, pp. 710 –720, Mar. 2007.

[17] IEEE 802.16e™ -2009, "IEEE Standard for Local and Metropolitan Area Networks – Part 16: Air Interface for Broadband Wireless Access Systems," May. 2009.

[18] L.-C. Wang and A. Chen, "Optimal Radio Resource Partition for Joint Contention- and Connection-Oriented Multichannel Access in OFDMA Systems," *IEEE Trans. Mobile Computing*, vol. 8, no. 2, Feb. 2009.
[19] C.-C. Wu et al., "An Empirical Evaluation of VoIP Playout Buffer Dimensioning in Skype Google Talk and MSN Messenger," In *Proc. ACM NOSSDAV*, 2009.
[20] C. Fraleigh, F. Tobagi, and C. Diot, "Provisioing IP Backbone Networks to Support Latency Sensitive Traffic," in *Proc. IEEE INFOCOM*, 2003.
[21] L. Ding and R. A. Goubran, "Speech Quality Prediction in VoIP using the Extended E-model," in *Proc. IEEE GLOBECOM*, Dec. 2003.
[22] Annex A: Silence Compression Scheme, ITU-T Rec. G.723.1, Int. Telecommun. Union, Nov. 1996.
[23] Coding of Speech at 8kbit/s using Conjugate-Structure Algebraic-Code-Excited Linear Prediction (CS-ACELP), ITU-T Rec. G. 729, Int. Telecommun. Union, Jan. 2007.
[24] Appendix II: A Comfort Noise Payload Definition for ITU-T G.711 Use in Packet-based Multimedia Communication Systems, ITU-T Rec. G.711, Int. Telecommun. Union, Feb. 2000.
[25] R. Srinivasan, J. Zhuang, L. Jalloul, R. Novak, and J. H. Park, "Draft IEEE 802.16m Evaluation Methodology Document," Apr. 2007.

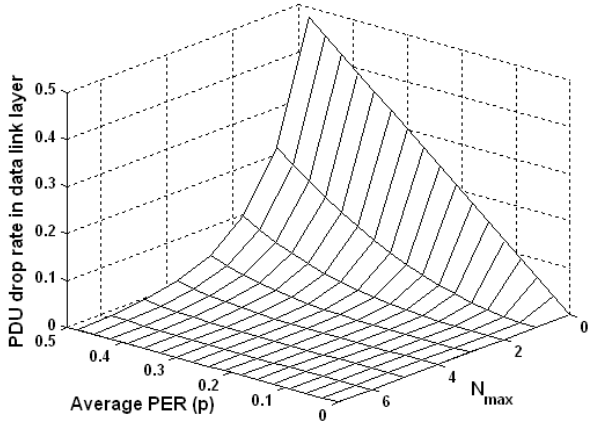

Figure 4. PDU drop rate of G.711 vs. average PER and N_{max}

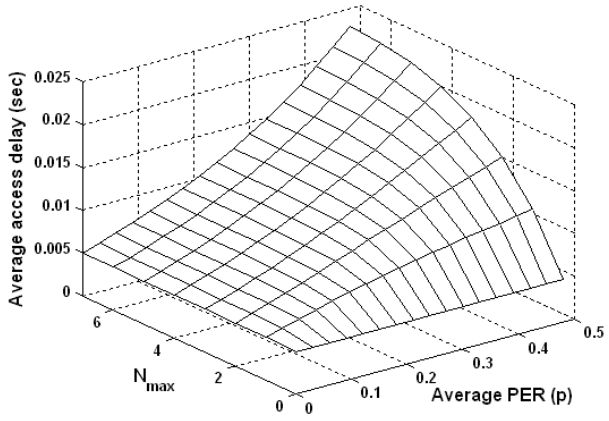

Figure 5. Average access delay (sec) of G.711 vs. average PER and N_{max}

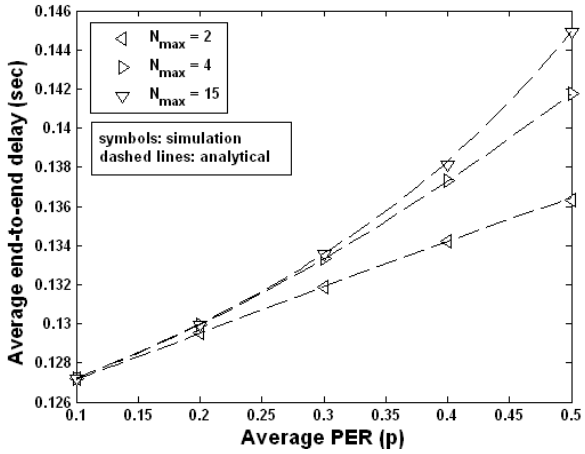

Figure 6. Average end-to-end delay of G.711 vs. average PER

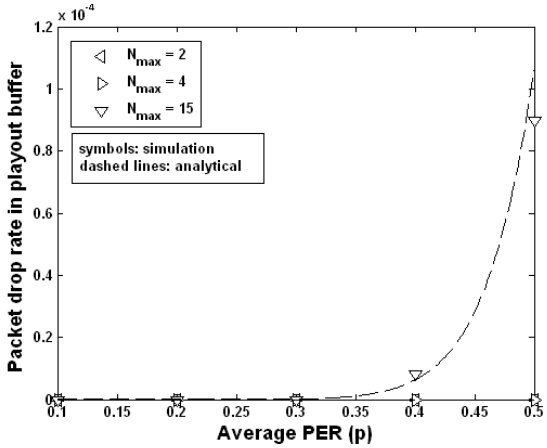

Figure 7. Packet drop rate of G.711 in the playout buffer vs. average PER

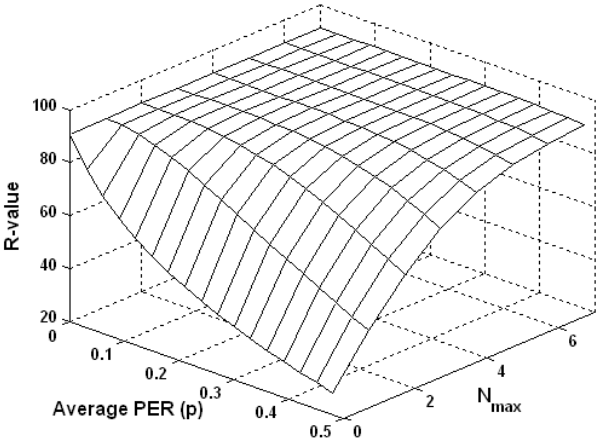

Figure 8. R-value of G.711 vs. average PER and N_{max}

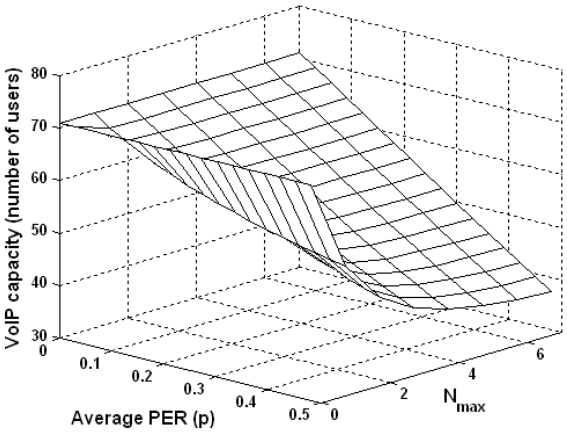

Figure 9. VoIP capacity of G.711 vs. average PER and N_{max}

TABLE 2. OPTIMAL N_{MAX} OF CROSS-LAYER OPTIMIZATION AND QOS REQUIREMENT BASED FOR VOIP SPEECH CODECS

VoIP Codec	Cross-layer Optimization				QoS Requirement Based (e.g. 50 msec)			
	R_{th}	$N^*_{\max}$	$\overline{m}(N^*_{\max})$	$\overline{R}(N^*_{\max})$	T_d	$N_{\max}$	$\overline{m}(N_{\max})$	$\overline{R}(N_{\max})$
G.711	70	1	59.59	70.46	20 msec	2	55.30	77.00
	80	3	53.26	85.41				
	90	5	51.75	90.02				
G.729	70	3	183.88	75.36	10 msec	4	180.4	78.27
	80	7	177.36	80.17				
	90	-	-	-				
G.723.1	70	4	338.59	70.53	30 msec	1	447.75	40.21
	80	-	-	-				
	90	-	-	-				