Smart Resource Allocation Algorithm Considering Voice Activity for VoIP Services in Mobile-WiMAX System

Howon Lee, Hyu-Dae Kim, and Dong-Ho Cho

Abstract—A new algorithm is proposed for wireless voice over IP (VoIP) codecs that allocates radio resources in consideration of the variations in packet size and packet-generation period. This algorithm allows resources to be utilized more efficiently in Mobile-WiMAX systems. The outstanding performance of our proposed algorithm with respect to average throughput and average packet dropping probability is demonstrated through intensive numerical analysis and simulations. VoIP capacity is compared with respect to the application of different VoIP codecs (G.729B, adaptive multirate (AMR)) and VoIP header compression schemes (payload header suppression (PHS) and robust header compression (ROHC)).

Index Terms—Mobile-WiMAX system, resource allocation, VoIP service, voice activity, payload header suppression, robust header compression.

I. INTRODUCTION

A. Reference System

MOBILE-WIMAX based on the IEEE 802.16e-2005 air-interface standard is a leading solution for wireless broadband services. This system is an improved version of the WiMAX system based on the IEEE 802.16-2004 standard, containing added features and attributes to support mobility [1]. Mobile-WiMAX can offer a high data rate, low latency, advanced security, quality of service (QoS), and low-cost deployment.

Voice over IP (VoIP) is a very important service for Mobile-WiMAX, because through VoIP technology, Mobile-WiMAX users can utilize voice services more cheaply compared with current mobile systems. Therefore, supporting as many voice users as possible while using limited radio resources is a major issue that could be a key to the success of the Mobile-WiMAX system. In this research area, the allocation of uplink resources is a central topic because most systems supporting VoIP are still uplink limited [2][3].

There are three uplink resource allocation algorithms for supporting VoIP services in the current Mobile-WiMAX system: unsolicited grant service (UGS), real-time polling service (rtPS), and extended-rtPS (ertPS) [4]. Among these algorithms, ertPS was proposed by us, and adopted in the IEEE 802.16e standard, in 2004 [5]. There are also a few related works on supporting VoIP services, such as UGS-AD in the DOCSIS system [6], UGS with GM-bit [7] and Oh’s algorithm [8]. However, these conventional algorithms have severe problems, such as waste of uplink resources, needless polling process, and additional functionality support [7][9]. Although ertPS was proposed recently, it also causes uplink resources to be wasted when VoIP codecs with variable packet-generation periods are used. In addition, in the case of Oh’s algorithm, even though this algorithm considers variations in the packet generation period by utilizing a random access scheme during inactive periods, the system needs new functionality to support the random access scheme for VoIP services. This could sometimes place a burden on the system and when the traffic load is very high, packets could collide, which can cause unpredictable VoIP packet delay. Eventually, it can give rise to packet drops.

In addition, R. Jain et al. presented key issues of system-level modeling for the IEEE 802.16e Mobile-WiMAX system [10] and F. Wang et al. evaluated the system performance in view of various features such as different types of MIMO scheme, receiver structures, and frequency reuse schemes [2]. In [11], M.-H. Fong et al. introduced a persistent resource allocation for enhancing the VoIP capacity based on dedicated allocations during “ON” duration. Moreover, Y. Choi et al. proposed a joint design of admission control and transmission rate adaptation method for the VoIP service [12].

B. VoIP Codecs for Mobile-WiMAX

Various VoIP codecs can be utilized for VoIP services in the Mobile-WiMAX system or the next-generation Mobile-WiMAX system. Each codec has different properties with respect to packet size, packet generation period, and the number of data rates. In this section, we introduce several candidates for VoIP codecs in the Mobile-WiMAX system [13][14].

1) Adaptive MultiRate (AMR): The frame duration of AMR is 20 ms. The codec has a total of nine data rates: eight voice-data generation periods (4.75 kbps ~ 12.20 kbps) and a silence descriptor (SID) generation period (1.80 kbps) [15].

2) Enhanced Variable Rate Codec (EVC): The frame duration of EVRC is 20 ms. The voice activity factor is 0.403 with 29% full rate (8.6 kbps), 4% half rate (4 kbps), 7% quarter rate (2 kbps), and 60% eighth rate (0.8 kbps). Full, half, and quarter rates are included in the ‘on’ duration.
(speech, active), and an eighth rate is included in the ‘off’ duration (silence, inactive). Case by case, eighth-rate frames could be blank (not transmitted) or not.

3) ITU-T G.729B: The frame duration of G.729B is 10 ms. This codec has two data rates, such as active (8 kbps) and inactive (0 kbps). If the transmission mode of the receiver is active, the receiver decodes speech data; otherwise, it generates comfort noise locally, on the basis of information about the level of background noise. The voice activity factor of G.729B is 0.4.

C. Organization of this paper

The remainder of this paper is organized as follows. In Section II, we introduce the conventional resource allocation algorithms for VoIP services and their problems. In Section III, we propose a smart resource allocation algorithm (ertPS+) that takes into account variations in packet size and packet-generation period for wireless VoIP codecs. In Section IV, we explain our resource allocation policy and analyze the performance of our proposed algorithm by using a two-state Markov-modulated Poisson process (MMPP) and discrete-time Markov chain (DTMC) based on a virtual uplink queue. In Section V, we demonstrate the superior performance of our proposed algorithm through numerical and simulation results. Section VI concludes this paper.

II. RELATED WORKS AND PROBLEMS

A. Dedicated Resource Allocation (UGS Algorithm)

The base station (BS) periodically assigns fixed-size grants to voice users [4]. These fixed-size grants are sufficient to send voice packets generated by the maximum data rate of VoIP codecs. However, in general, voice users do not always have the same size voice-packets, because they have variable data rates, as described in Section I-B. So, in UGS, a large number of uplink resources are wasted because the BS always allocates the same amount of uplink resources to each user, regardless of his voice status [9].

B. Polling-Based Resource Allocation (rtPS Algorithm)

In this algorithm, the BS assigns uplink resources that are sufficient for unicast bandwidth requests to the voice user [4]. Due to the fact that this algorithm always uses a bandwidth request process for suitable size grants, it transports data more efficiently than UGS. However, this bandwidth request process always causes polling overhead and additional access delay [9]. In order to avoid the polling process in the ‘off’ period, a minimum polling size could be the size of the voice packet generated by the minimum data rate of the VoIP codec. In this case, there is no polling overhead in the ‘off’ period. However, depending on the VoIP codec, this assumption may result in a waste of uplink resources. So, the minimum polling size must be determined carefully to maximize system performance.

C. Hybrid Resource Allocation Algorithm

1) UGS-AD algorithm: This algorithm has two allocation modes (UGS and rtPS) and can switch between these modes according to the status of voice users [6]. When the data rate of a user changes to the minimum rate, the user notifies his status by utilizing piggyback-based bandwidth requests of zero bytes. Then, the BS switches its mode from UGS to rtPS. By contrast, when the data rate of the voice user increases to the full rate, the user should transmit unsolicited bandwidth requests of non-zero bytes. In this case, at first, the BS allocates sufficient uplink resources to the user to send the first delayed voice packet and a second full-rate packet that is generated by VoIP codecs. Then, the BS periodically assigns uplink resources according to the general operation of the UGS mode.

UGS-AD partially can solve the problems caused by UGS and rtPS, provided that the voice users use VoIP codecs that have two data-rates and no variations in the packet-generation period. However, in the case of VoIP codecs with variable data rates, this algorithm cannot solve the problems completely [9]. For example, in the case of half and quarter rates in EVRC, uplink resources would be wasted, because the BS maintains UGS mode. In addition, when there are variations in the packet-generation period in the VoIP codecs, this algorithm wastes lots of resources in inactive periods.

2) UGS with GM-bit (Lee’s algorithm): Lee’s algorithm can partially solve the problems of UGS and rtPS, in that the BS basically assigns uplink resources to voice users by considering only their on-off transitions [7]. This algorithm uses one reserved bit in the conventional generic MAC header to inform the BS of the user’s voice state transitions [4]. This reserved bit is defined as a Grant-Me (GM) bit in this algorithm. When the voice state of the user is ‘on’, the GM bit is set to ‘1’. Then the BS assigns the maximum grant size that is sufficient to send voice packets. Otherwise, the GM bit is set to ‘0’. Then, the BS assigns the minimum grant size that is sufficient to inform the BS of the user's voice state transitions.

3) Extended-rtPS algorithm: When the voice data rate is decreased, the user requests reduced bandwidth using the extended piggyback request bits of the GMSH (grant management subheader) [4]. There is no waste, because the user utilizes the remaining uplink resources assigned to him. In ertPS, the BS assigns uplink resources according to the requested size periodically, until the voice user requests another size of bandwidth. On the contrary, when the voice data rate is increased, the user requests increased bandwidth using the bandwidth request bits of the BUH (bandwidth request and uplink (UL) transmission (Tx) power report header) [4]. Then, the BS performs the first bandwidth allocation in the next MAC frame following the request process. The second allocation of bandwidth is made after the elapse time of ‘the basic polling period minus the MAC frame duration’ [5]. After that, the BS assigns uplink resources periodically, according to the requested size.

When VoIP codecs that have variations in packet-size and packet-generation period are used, uplink resources might be wasted in the ertPS algorithm, because ertPS does not consider variation in the period of packet generation.

III. PROPOSED ALGORITHM (EXTENDED-rtPS+)

To support lots of VoIP users, we should consider the following two characteristics of VoIP codecs: variations in
VoIP packet-size and VoIP packet-generation period. In this section, we propose a new algorithm for allocating uplink resources (ertPS⁺), taking into account these characteristics of VoIP codecs (EVRC, G.729B and AMR) for the Mobile-WiMAX system. For clarity, the notations used in this paper are summarized in Table I. A.

Consideration of the Known Packet-Generation Period in Inactive Periods

With AMR, apart from the SID-FIRST frame, all SID-UPDATE frames are generated every 8th frame in the SID generation period [16]. In this case, if we allocate uplink resources according to the general packet-generation period (Tbasic), lots of uplink resources may be wasted. So, in the ertPS⁺ algorithm, through piggybacks or unsolicited indications of the next packet-generation period, such as the procedures of bandwidth requests in ertPS, we can easily solve the problem of wasted resources during inactive periods.

If the VoIP codec enters the SID generation state, users can be aware of their status because the size of SID packets is different from the sizes of other voice packets. Thus, the VoIP user can easily inform the BS of his altered packet size and packet-generation period. Then, the BS allocates uplink resources periodically according to the requested packet size and packet-generation period, as shown in Fig. 1. However, because VoIP services are very sensitive to delays, there is a limit on how much the BS can increase the period of resource allocation. Hence, a maximum allocation period (Tmax) should be determined, in order to satisfy the VoIP delay bound (Tbound). Here, Tmax (≤Tbound) could be chosen adaptively according to whether the interworking system is mobile-to-mobile or mobile-to-land. For example, we can set this value to 40ms or 60ms for the mobile-to-mobile environment, considering such matters as maximum end-to-end delay bound (285ms by ITU-T), backbone delay, and packet processing delay [17]. If the environment is mobile-to-land, since the VoIP delay bound could be more relaxed, we can set Tmax as a larger value compared with the case of the mobile-to-mobile.

In the SID-generation period, if the resources that are required for the SID-packet transmission are much larger than those required for transmission of the BUH, the BS can periodically allocate uplink resources just for the header transmissions to prevent uplink resources from being wasted. By contrast, if the resources that are required for SID packet transmission are smaller than, or nearly the same as, those required for the header transmissions, the BS should allocate resources according to the resources that are required for the SID-packet transmissions.

After the SID-generation period, the VoIP users cannot send their voice packets by using uplink resources that are assigned for the transmission of SID packets. This is because voice packets are generally larger than SID packets. The VoIP users can transmit only the BUH using these resources. Moreover, due to the fact that the allocation period of the SID packets is larger than that for voice packets, the VoIP users may have some packets that are not transmitted, but queued within the delay bound. However, this problem can be solved because the VoIP user requests his required bandwidth for the sum of the sizes of the queuing packets. In this case, the BS allocates all of the requested bandwidth in the next MAC frame. Then, following the bandwidth allocation is based on the initial time of the bandwidth request process, and is made immediately after the basic polling period, as shown in Fig. 1.

To apply ertPS⁺ to Mobile-WiMAX, it is necessary to make some modifications to the extended piggyback request field of the GMSH and the bandwidth request field of the BUH. In the case of VoIP services, it may be possible to reduce
and modify the size of these fields, because voice packets are generally smaller than other types of data packets. In the IEEE 802.16e-2005 system, the extended piggyback request field of the GMSH and the bandwidth request field of the BUH have the same size of 11 bits, respectively. In order to apply ertPS+, the field for indication of packet-generation period should be included. So, by modifying a few bits (e.g., 3~4 bits) of bandwidth request field of each header, we can simply apply our proposed algorithm to the Mobile-WiMAX system.

In ertPS and ertPS+, we can use codeword transmissions over a channel quality information channel (CQICH) for bandwidth requests. For example, if the user has a CQICH, the BS does not need to assign uplink resources periodically during the ‘off’ period in which VoIP packets are blanked. When the VoIP codec enters the ‘on’ period, the user can easily request bandwidth through the CQICH. However, CQICH-based requests are not normally used for VoIP services, because it is not practical to allocate a CQICH to all VoIP users. Namely, it would generate a large MAC overhead and lots of uplink resources would be wasted. In addition, even if the CQICH codeword of the Mobile-WiMAX system were used, resources would still be wasted because there is no codeword that is suitable for the transmission of SID packets. In the current Mobile-WiMAX system, there already exists a CQICH codeword for the ertPS algorithm. However, this codeword is just designed for sending maximum-size packets that are generated in the maximum data rate of the VoIP codec. Therefore, we do not here address the case in which the VoIP user has a CQICH.

B. Consideration of the Unknown Period of Packet Generation During Inactive Periods

In the case of G.729B or EVRC whose VoIP packets of eighth data-rate are blanked, voice packets are not generated during an inactive period and we cannot estimate the duration of the inactive period. Hence, if we use the conventional resource allocation algorithms, which do not consider the variation in the packet-generation period, many uplink resources could be wasted. To solve these problems, we propose a maximum increment strategy.

Then, the resource-allocation period is adjusted as follows:

$$T_{alloc} = \begin{cases} T_{max}, & \text{on inactive duration} \\ T_{basic}, & \text{on active duration} \end{cases}$$

In equation (1), $T_{basic}$ is a basic packet-generation period of the VoIP codec in active mode. When the VoIP codec enters inactive mode, the user notifies his status to the BS. As shown in Fig. 2, there is no overhead for notifying his status because he uses resources that are allocated for the transmission of voice packets. In this case, the user can directly notify his next allocation time, or the BS can change $T_{alloc}$ into $T_{max}$ adaptively, by considering the user’s voice status. In addition, as described in section III-A, we cannot increase $T_{max}$ boundlessly. We should choose a proper value within the delay bound that can satisfy the quality of VoIP services.

When the VoIP codec enters active mode from inactive mode, resources are allocated based on the same way as the packet-generation period is known. Fig. 2 shows the overall operation of ertPS+ for the maximum increment strategy when the packet-generation period is unknown.

IV. RESOURCE ALLOCATION POLICY AND PERFORMANCE ANALYSIS

A. Overall Resource Allocation Policy

In uplink radio resource allocation, we should consider two kinds of allocation:

1) Resource allocation for users who are in the off-state, and
2) Resource allocation for users who are in the on-state.

In case of users in the off-state, we allocate uplink resources by taking into account the maximum allocation period ($T_{max}$) of the proposed resource allocation algorithm (ErtPS+), which is based on the VoIP delay bound ($T_{bound}$), as described in Section III. In the case of users in the on-state, we assign uplink resources according to the packet-generation period of the VoIP codec, because the VoIP packets of the users are generated periodically.

In this section, in order to analyze the performance of our proposed scheme, we assume the VoIP traffic model to be a two-state MMPP model and formulate the system as a DTMC that is based on a virtual uplink queue at the BS [18][19].

B. Resource Allocation for Users in Off-state

The BS should assign uplink resources to users who are in the off-state, because these users might have packets to transmit (e.g., SID) or the BS does not know when the active period starts. In this paper, we call these resources that are allocated to VoIP users who are in the off-state polling resources. The probability that the status of the users is inactive (= off) can be obtained by $p_{off} = \frac{\beta^{-1}}{\alpha^{-1} + \beta^{-1}}$. Here, $1/\alpha$ and $1/\beta$ are the mean values of the on and off periods, which are distributed exponentially [20]. Given that we should periodically allocate polling resources to users in off-state every $T_{max}$ ms, the conditional probability that the polling resources are given among the off-users ($p_{off, act}$) can be calculated as

$$p_{off, act} = Pr\{\text{user receives polling resources} | \text{VoIP status is inactive}\} = \frac{T_{MAC}}{T_{max}}.$$
Here, $T_{MAC}$ is a MAC frame duration in the Mobile-WiMAX system. By using $p_{oфф}$ and $p_{oфф,act}$, we can calculate the probability that users receive polling resources ($p_{poll}$), which is described as follows:

$$p_{poll} = p_{oфф} \times p_{oфф,act} = \frac{\beta^{-1} \cdot T_{MAC}}{(\alpha^{-1} + \beta^{-1}) \cdot T_{max}}.$$  \hspace{1cm} (3)

Then, we can obtain the resource allocation set for off-state users ($\psi_{oфф}$). The average size of $\psi_{oфф}$ is calculated by $N_{avg}(\psi_{oфф}) = p_{poll} \times N$. The amount of resources that is required to allocate to users who are in the off-state ($R_{oфф}$) at a certain MAC frame is given by

$$R_{oфф} = \sum_{m} x_{m,off} \times I_{m,poll}.$$  \hspace{1cm} (4)

In equation (4), $I_{m,poll}$ is the required number of uplink resources for polling when the VoIP user utilizes the m-th MCS level, and $x_{m,off}$ is the number of users who use m-th MCS level in the resource allocation set for off-users ($\psi_{oфф}$). So, through our resource allocation policy, the total amount of available resources for users who are in the on-state is $R_{on} = R_{tot} - R_{oфф}$.

### C. Resource Allocation for Users in On-state

1) Traffic Model: In general, we can formulate the VoIP traffic generated by a single VoIP user as a simple on-off model [20]. Furthermore, all traffic generated by the VoIP users in the cell can be modeled as a two-state MMPP model [18][21], as shown in Fig. 3. The MMPP is a stochastic process in which the intensity of a Poisson process is defined by the states of a Markov chain. That is, the Poisson process can be modulated by the Markov chain. MMPP is a special case of the Markovian arrival process (MAP). This model is very suitable for formulating the multi-user VoIP traffic, because the MMPP captures the interframe dependency between consecutive frames. The transition rate matrix and the Poisson arrival rate matrix of the MMPP in Fig. 3 can be expressed as follows:

$$R = \begin{bmatrix} -r_1 & r_1 \\ r_2 & -r_2 \end{bmatrix}, \quad A = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}.$$  \hspace{1cm} (5)

In order to utilize the MMPP model, we should match the MMPP parameters in equation (5) with the parameters of the simple on-off model. Several methods have been proposed for matching the parameters of $r_1$, $r_2$, $\lambda_1$ and $\lambda_2$ in the MMPP model, such as the moment-based matching technique, the $\Sigma$-matching technique, and the index of dispersion for counts (IDC) matching technique [22]-[25]. We here adopt the IDC matching technique, because it yields adequate results for the matching of parameters and has appropriate computation complexity compared with other matching techniques. Then, $r_1$, $r_2$, $\lambda_1$, and $\lambda_2$ in the equation (5) can be calculated by

$$r_1 = \frac{2(\lambda_2 - \lambda_{avg})(\lambda_{avg} - \lambda_1)^2}{(\lambda_2 - \lambda_1)\lambda_{avg}(IDC(\infty) - 1)}. \hspace{1cm} (6)$$

$$r_2 = \frac{2(\lambda_2 - \lambda_{avg})^2(\lambda_{avg} - \lambda_1)}{(\lambda_2 - \lambda_1)\lambda_{avg}(IDC(\infty) - 1)}. \hspace{1cm} (7)$$

$$\lambda_1 = A \cdot \frac{\sum_{i=0}^{N_{act,avg}} i \cdot \pi_i}{\sum_{j=0}^{N_{act,avg}} \pi_j}, \quad \lambda_2 = A \cdot \frac{\sum_{i=N_{act,avg}+1}^{N} i \cdot \pi_i}{\sum_{j=N_{act,avg}+1}^{N} \pi_j}. \hspace{1cm} (8)$$

Here, $N$ is the total number of VoIP users in the system and $A$ is the emission rate in the on-state $(A = \frac{1}{T_{basic}})$. The average arrival rate in the simple on-off model is $\lambda_{avg} = N \times A \times p_{on}$. The probability that the status of the users is active can be obtained by $p_{on} = \frac{\alpha^{-1}}{\alpha^{-1} + \beta^{-1}}$. Then, the average number of active users is $N_{act,avg} = [N \times p_{on}]$ and the steady-state probability of a one-dimensional Markov chain when considering $N$ independent simple on-off voice users can be calculated by $p_i = \binom{N}{i} p_{on}^i (1 - p_{on})^{N-i}$. In equation (6) and (7), $IDC(\infty)$ is given as [21]

$$IDC(\infty) = 1 + \frac{2(\lambda_1 - \lambda_2)^2 r_1 r_2}{(r_1 + r_2)^2(\lambda_1 r_2 - \lambda_2 r_1)}. \hspace{1cm} (9)$$

In Fig. 3, ‘UL’ and ‘OL’ denote underloading and overloading states, $\lambda_1$ and $\lambda_2$ are the packet arrival rates in the underloading and overloading states, respectively, and ‘$r_1$’ and ‘$r_2$’ represent the transition rates between the underloading and overloading states of the two-state MMPP model, respectively. In general, the packet arrival rate ($\lambda_i$) is determined by phase $i$ of the Markov chain and the total number of phases in the MMPP model.
2) Throughput and Packet Dropping Probability: The behavior of queuing packets in a virtual uplink queue can be analyzed with a DTMC model, as shown in Fig. 4. In this figure, each state denotes the number of VoIP packets that are queued in the virtual uplink queue. In this model, the transition matrix \( P \) of the virtual uplink queue can be expressed as:

\[
P = \begin{bmatrix}
    p_{0,0} & p_{0,1} & p_{0,2} & \cdots & p_{0,L_{max}} \\
p_{1,0} & p_{1,1} & p_{1,2} & \cdots & p_{1,L_{max}} \\
p_{2,0} & p_{2,1} & p_{2,2} & \cdots & p_{2,L_{max}} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
p_{L_{max},0} & p_{L_{max},1} & p_{L_{max},2} & \cdots & p_{L_{max},L_{max}}
\end{bmatrix}
\]

(10)

Here, \( L_{max} \) is the maximum size of the virtual queue. In this matrix (\( P \)), each element \( (p_{i,j}) \) is a 2-by-2 matrix and represents the transition of the number of VoIP packets in the virtual uplink queue. In other words, there are \( i \) queuing packets in the current MAC frame, and there will be \( j \) queuing packets in the next MAC frame. When the number of queuing packets is \( i \), given that \( k \) packets are scheduled, \((j - \max(i - k, 0))\) should arrive so that the number of packets in the virtual queue becomes \( j \). Then, each element of \( P \) \((p_{i,j})\) can be represented as follows:

\[
p_{i,j} = \sum_{k=0}^{\min(i,j)} \Pr\{k \text{ packets are serviced}\} \cdot \Pr\{j + k - i \text{ packets arrive}\}
\]

\[
= \sum_{k=0}^{\min(i,j)} \Pr\{k \text{ packets are scheduled}\} \cdot \Pr\{j + k - i \text{ packets arrive}\}
\]

\[
+ \sum_{k=\max(i,j)}^{\infty} \Pr\{k \text{ packets are scheduled}\} \cdot \Pr\{j \text{ packets arrive}\}
\]

\[
= \sum_{k=0}^{\max(i,j)} \Pr\{k \text{ packets are scheduled}\} \cdot \Pr\{N(\psi_{on}) = k\} \cdot U(j - \max(i - k, 0)).
\]

(11)

In equation (11), \( R_{basic,max} \) is the number of uplink resources that are required to transmit a VoIP packet that is generated in the on-state when we use the highest MCS level (i.e., 64QAM and 3/4 coding in the Mobile-WiMAX), and \( \psi_{on} \) is the resource allocation set for on-state users. Throughout the summation for the possible number of packets which can be scheduled during the MAC frame, we can calculate the 2-by-2 transition probability matrix that represents the transition of the number of VoIP packets in the virtual uplink queue.

In the two-state MMPP model, the transition probability matrix \( (U) \) and the diagonal probability matrix can be obtained as follows:

\[
U = (\Lambda - R)^{-1} \Lambda
\]

(12)

\[
D(m) = \\
\begin{bmatrix}
    (\lambda_1 T_{MAC})^m e^{(-\lambda_1 T_{MAC})} & 0 \\
    0 & (\lambda_2 T_{MAC})^m e^{(-\lambda_2 T_{MAC})}
\end{bmatrix}
\]

(13)

In equation (13), each element of \( D(m) \) is the probability that \( m \) VoIP packets arrive at the BS for the MAC frame duration \( (T_{MAC}) \) in each phase of the two-state MMPP. In addition, in equation (11), \( \Pr\{N(\psi_{on}) = k\} = P_i(k) \) is the probability that the BS schedules \( k \) VoIP packets from the virtual uplink queue. This probability may vary according to the distribution of MCS levels for VoIP users, which can be obtained as:

\[
P_i(k) = \frac{M(k)}{\sum_{l=0}^{R_{basic,max}} M(l)}
\]

(14)

Here, \( M(k) \) is the number of cases that \( k \) packets are scheduled during a MAC frame from the virtual uplink queue.

Through the transition matrix \( (P) \) in (10), we can obtain the steady-state probability matrix for the two-state MMPP model \( (\pi_{MMPP}) \). This matrix can be calculated by solving equations \( \pi_{MMPP} \cdot P = \pi_{MMPP} \) and \( \pi_{MMPP} \cdot 1 = 1 \). Here, \( \pi_{MMPP} \) is 1 by \( (L_{max} + 1) \) matrix. So, the probability that \( k \) VoIP packets are queued in the virtual uplink queue \( (\pi(k)) \) can be calculated as:

\[
\pi(k) = \pi_{MMPP}(2k) + \pi_{MMPP}(2k + 1).
\]

(15)

Through equation (15), the steady-state probability matrix of the DTMC model \( (\pi) \) in Fig. 4 can be obtained as:

\[
\pi = [\pi(0) \; \pi(1) \; \pi(2) \; \cdots \; \pi(L_{max})].
\]

(16)

So, by using equation (16), we can obtain various performance results, such as average arrival rate, the average length of the virtual uplink queue, the average number of serviced VoIP packets, average VoIP uplink throughput, and the probability that VoIP packets will be dropped according to the number of VoIP users. First, the average queue length \( (L_{avg}) \) and the average arrival rate in the two-state MMPP model \( (\rho) \) can be calculated by:

\[
L_{avg} = \sum_{i=0}^{L_{max}} i \cdot \pi(k).
\]

(17)

\[
\rho = s \cdot \sum_{m=0}^{N \cdot A_{max}} m \cdot D(m) \cdot 1.
\]

(18)

In equation (18), \( s \) is calculated by solving \( s \cdot U = s \), and \( 1 \) is a column matrix of ones. \( A_{max} \) is the maximum number of packets that can arrive from each VoIP user during \( T_{MAC} \). Similarly, the average number of serviced VoIP packets \( (N_{avg}(\psi_{on})) \) can be obtained as follows:

\[
N_{avg}(\psi_{on}) = \frac{R_{safe} - R_{aff}}{R_{safe} - R_{aff}} \sum_{i=0}^{L_{max}} \sum_{j=0}^{k_{max}} \min(i, j) \cdot \pi(j) \cdot \Pr\{N(\psi_{on}) = i\}.
\]

(19)

Here, given that we have already allocated uplink resources \( (R_{aff}) \) for the VoIP users in the off-state \( (\psi_{off}) \), we can simply assign \( R_{safe} - R_{aff} \) uplink resources to the users. Then, the average VoIP uplink throughput for \( \psi_{on} \) is represented as:

\[
S_{avg} = N_{avg}(\psi_{on}) \cdot l_{PDU}.
\]

(20)

In equation (20), \( l_{PDU} \) is the size of VoIP PDU. In addition, we can calculate the dropping probability \( (P_{drop}) \) of VoIP packets as follows:

\[
P_{drop} = \frac{1 - N_{avg}(\psi_{on})}{\rho}.
\]

(21)
Hence, by using equation (21), we obtain the VoIP capacity as follows:

\[
C_{V,oP} = \arg\max N \in \{N \mid P_{\text{drop}} \leq P_{\text{limit}}\} = \arg\max N \in \left\{N \mid \frac{\rho - N_{\text{avg}}(\psi_{\text{on}})}{\rho} \leq P_{\text{limit}}\right\} \tag{22}
\]

Here, \(P_{\text{limit}}\) is the threshold of the packet dropping probability for VoIP services.

V. NUMERICAL AND SIMULATION RESULTS

A. Analysis and Simulation Environments

We evaluated the uplink VoIP performance of Mobile-WiMax. The duration of a MAC frame was 5 ms, and it consisted of 48 symbols and 1024 OFDM-subcarriers (840 data and pilot subcarriers). Here, the first symbol is used for a preamble. The ratio of the symbols of the downlink frame to those of the uplink frame was 29:18. In the uplink, given that three symbols are usually used for uplink control signaling, we used the remaining 15 symbols for voice-packet transmissions. Although the remaining symbols could be used for diversity and band-AMC transmissions, we assumed that these symbols are used for VoIP packet transmissions.

To obtain the numerical and simulation results, we used the following distribution for modulation and coding (MCS) levels: quaternary phase shift keying (QPSK) 1/12 = 3.71%, QPSK 1/8 = 12.01%, QPSK 1/4 = 29.10%, QPSK 1/2 = 29.67%, QPSK 3/4 = 9.23%, 16-quadrature amplitude modulation (QAM) 1/2 = 12.51%, 64-QAM 1/2 = 0.75%, and 64-QAM 3/4 = 3.02% [26]. This distribution was obtained when the users were moving at 120km/h. The full size of the real-time protocol (RTP)/user datagram protocol (UDP)/internet protocol version 4/version 6 (IPv4/IPv6) header is 55 bytes, and it is compressed by payload header suppression (PHS) and robust header compression (ROHC), were 16 and 2 bytes, respectively.

We mainly utilized the G.729B codec in the simulation and applied two different header compression methods: PHS and ROHC. In addition, VoIP services are very delay-sensitive; hence, it is difficult to apply resource allocation (scheduling) policies, such as the MAX C/I scheduler or the proportional fair (PF) scheduler, for the purpose of maximizing capacity. Thus, a round-robin (RR) scheduler is generally utilized for VoIP services, because it can guarantee fairness to VoIP users, and provide constant delay and jitter performance for VoIP users [27]. Therefore, in this paper, we used a RR scheduler.

The assumptions in the simulations are summarized as follows:

- MAP overhead was not taken into account.
- \(T_{\text{bound}}\) was 60 ms considering the maximum end-to-end delay, and \(T_{\text{max}}\) was 40 ms.
- VoIP packets of which delays are greater than \(T_{\text{bound}}\) were discarded.
- A utilized service was only VoIP.
- A round robin scheduler was utilized.
- A voice activity detector operated perfectly.
- VoIP users did not use polling resources to transmit their VoIP packets.

B. Numerical and Simulation Results

In Fig. 5, ‘Arrived’ means VoIP packets generated by the MMPP traffic model in the case of analysis, and the total number of packets generated by each simple on-off model of VoIP users in the case of simulation, respectively. When the number of VoIP users increases, the number of ‘Arrived’ packets increases linearly. By contrast, in the case of ‘Serviced’ packets, the number of ‘Serviced’ packets increases according to the increment of VoIP users up to a certain number. We here define this certain number as a threshold number. However, above the threshold number, the number of ‘Serviced’ packets does not increase further. Since the total number of resources is limited, resource saturation occurs. Due to the influence of resource allocations for users in the off-state, if the total number of users increases, the available amount of resources for users in the on-state is reduced. Hence, after resource saturation, the numbers of ‘Serviced’ packets of ertPS and ertPS+ decrease. Due to the fact that our proposed resource allocation algorithm (ertPS+) allocates fewer resources to inactive users than ertPS in the Mobile-WiMAX system, the diminution rate of ertPS+ is much smaller than that of ertPS. Because of this saving, the number of VoIP packets that are serviced in the proposed algorithm is greater than that in ertPS. In the case of ertPS, regardless of the packet generation-periods, the BS allocates resources to users periodically. Then, the problem of resource waste becomes severe. Thus, the saving of resources is one of the main contributions of our proposed algorithm.

Fig. 6 shows the average throughput against the number of VoIP users. The average throughput increases linearly according to the increment in the number of VoIP users, up to the threshold number. Since resource saturation occurs at the threshold number, the BS cannot assign uplink resources to surplus users beyond the threshold number. Thus, even though the number of users increases, the average throughput cannot be increased without limit. After resource saturation occurs, the average throughput decreases slightly according to the increment of the number of VoIP users. This is because the number of required resources for users in the off-state increases according to the increment in the total number of users. From Fig. 7, we can see the number of resources that are required for users who are in the off-state against the increment in the number of voice users. In general, resource allocations for users in the off-state are not related to the utilized header compression methods, because the BS just allocates an adequate amount of resources for sending the BUH. Thus, in the case of the ROHC and PHS, the required resources for users who are in the off-state are the same. In summary, when the number of users in the off-state increases, the total amount of polling resources \(R_{\text{off}}\) increases. Thus, \(R_{\text{on}}\) decreases.

Since the resource utilization efficiency is greater in our proposed algorithm than in the ertPS, our proposed algorithm has higher average throughput and the decrement rate is smaller compared with the ertPS algorithm. In addition, from Fig. 6, it can be seen that the performance of the analysis and the simulation is almost the same.

Fig. 8 shows the VoIP packet dropping probability against...
VoIP users compared with PHS.

To demonstrate additional performance excellency of our proposed scheme, we have added simulation results obtained when we utilize a AMR codec, as shown in Fig. 9. The used data-rates of the AMR codec for simulations are 7.95 and 1.75 (SID) kbps, and the packet generation periods of voice-data, SID-FIRST, and SID-UPDATE are 20, 60, and 160 ms, respectively \[16\]. From Fig. 9, we can still obtain better performance in the case of using the erTPS+ algorithm compared with the erTPS algorithm even when we utilize the AMR codec. As mentioned before, if we use a VoIP codec which has the characteristics of variable packet-sizes and packet-generation periods, we always can improve the VoIP capacity compared with the conventional algorithms by adopting the erTPS+ algorithm. Since the AMR codec also has the variations of packet-size and packet-generation period, we can save the wasted resources in the erTPS algorithm, especially during the SID period.

Moreover, Fig. 10 and Fig. 11 demonstrate the packet dropping probability according to the number of voice users when control-packet errors occur. Here, the control-packets

the number of VoIP users. Here, we assume that the threshold of the packet dropping probability is 2%. The numerical results and simulation results are almost the same. We find that when we utilize the PHS method, erTPS and our proposed algorithm can support 65 and 80 VoIP users, respectively. When we use the ROHC method, erTPS and our proposed algorithm can support 93 and 122 VoIP users, respectively. According to the header compression method, erTPS+ has capacity gains of 23.1% and 31.2% when using PHS and ROHC, respectively. These gains are achieved principally by the allocation of resources to users who are in the off-state. By adaptation of the polling duration, lots of uplink resources can be saved compared with the conventional algorithm. As mentioned above, if we compress the upper-layer header through ROHC, since the VoIP PDU compressed by ROHC size is smaller than that compressed by PHS, we can support more users compared with the PHS method, as shown in Fig. 8. That is, since the compressed size of the RTP/UDP/IP header in ROHC (2 bytes) is smaller than the header size of PHS (16 bytes), in case that we use ROHC, we can support 60% more

VoIP users compared with PHS.

To demonstrate additional performance excellency of our proposed scheme, we have added simulation results obtained when we utilize an AMR codec, as shown in Fig. 9. The used data-rates of the AMR codec for simulations are 7.95 and 1.75 (SID) kbps, and the packet generation periods of voice-data, SID-FIRST, and SID-UPDATE are 20, 60, and 160 ms, respectively \[16\]. From Fig. 9, we can still obtain better performance in the case of using the erTPS+ algorithm compared with the erTPS algorithm even when we utilize the AMR codec. As mentioned before, if we use a VoIP codec which has the characteristics of variable packet-sizes and packet-generation periods, we always can improve the VoIP capacity compared with the conventional algorithms by adopting the erTPS+ algorithm. Since the AMR codec also has the variations of packet-size and packet-generation period, we can save the wasted resources in the erTPS algorithm, especially during the SID period.

Moreover, Fig. 10 and Fig. 11 demonstrate the packet dropping probability according to the number of voice users when control-packet errors occur. Here, the control-packets
denote BUH and GMSH. Compared with the ertPS algorithm, since our proposed algorithm should control even the resource allocation period, it is more sensitive to the control-packet errors than the ertPS algorithm. Especially, when the voice-status of the user is changed from “ON” to “OFF” (“ON-OFF”), the control-packet errors can cause more waste of uplink resources in the ertPS+ algorithm because the BS still allocates the resources in accordance with the frame duration of voice-data packets. On the other hand, in the case of a “OFF-ON” transition, just more voice packets would be queued. Thus, it causes negligible performance degradation. Here, some packets whose delays are greater than \( T_{\text{bound}} \) are dropped. Consequently, even though control-packet errors cause more packet drops compared with the conventional algorithm, we can show that the performance degradation with respect to the packet dropping probability is not severe even when probability of control-packet errors \( (P_{\text{ctl}}) \) is 0.1. In practical, a frame error-rate in the MAC layer is less than 0.01. Therefore, we can conclude that our proposed algorithm always has superior performance regardless of data-packet and control-packet errors.

VI. CONCLUSIONS

There are many VoIP codecs for wireless VoIP services, with different features and attributes. However, the algorithms for allocating resources that have been proposed previously have not considered the variations of packet size and the period of packet-generation. Therefore, we proposed an uplink resource allocation algorithm \( (\text{ertPS}^+) \) for these VoIP codecs, in order to maximize the VoIP capacity of the Mobile-WiMAX system. Through intensive analysis and simulation, we demonstrated the superior performance of our proposed algorithm. Our proposed algorithm can improve VoIP capacity by 23.1\% and 31.2\%, respectively, when we use G.729B with PHS and G.729B with ROHC, compared with the conventional ertPS algorithm. Moreover, when channel errors arise or we utilize another VoIP codec, such as the AMR codec, the VoIP capacity of our proposed scheme is always larger than the conventional algorithm. In the future, we will extend our analysis in the consideration of both queue and channel states.

REFERENCES


