This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

Improving Quality of VoIP Streams over WiMax

Shamik Sengupta, Mainak Chatterjee and Samrat Ganguly

Abstract

Real-time services such as VoIP are becoming popular and are major revenue earners for network service providers. These services are no longer confined to the wired domain and are being extended over wireless networks. Though some of the existing wireless technologies can support some low bandwidth applications, the bandwidth demands of many multimedia applications exceed the capacity of these technologies. IEEE 802.16 based WiMax promises to be one of wireless access technologies capable of supporting very high bandwidth applications.

In this paper, we exploit the rich set of flexible features offered at the medium access control (MAC) layer of WiMax for construction and transmission of *MAC protocol data units* (MPDU) for supporting multiple VoIP streams. We study the quality of VoIP calls, usually given by R-score, with respect to delay and loss of packets. We observe that loss is more sensitive than delay, hence we compromise the delay performance within acceptable limits in order to achieve a lower packet loss rate. Through a combination of techniques like forward error correction, automatic repeat request, MPDU aggregation, and minislot allocation, we strike a balance between the desired delay and loss. Simulation experiments are conducted to test the performance of the proposed mechanisms. We assume a three-state Markovian channel model and study the performance with and without retransmissions. We show that the feedback-based technique coupled with retransmissions, aggregation, and variable length MPDUs are effective and increases the R-score and mean opinion score by about 40%.

Keywords: VoIP, R-score, WiMax, FEC, ARQ, aggregation, fragmentation.

S. Sengupta and M. Chatterjee are with the School of Electrical Engineering and Computer Science, University of Central Florida, Orlando, FL 32816. Email: {shamik,mainak}@eecs.ucf.edu. S. Ganguly is with NEC Laboratories America, Princeton, NJ 08540. Email: samrat@nec-labs.com.

Digital Object Indentifier 10.1109/TC.2007.70804

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

I. Introduction

In-spite of the growing popularity of data services, voice services still remain the major revenue earner for the network service providers. The two most popular ways of providing voice services are the packet switched telephone networks (PSTN) and the wireless cellular networks. Deployment of both these forms of networks require infrastructures that are usually very expensive. Alternative solutions are being sought that can deliver good quality voice services at a relatively lower cost. One way to achieve low cost is to use the already existing IP infrastructure. Protocols used to carry voice signals over the IP network are commonly referred to voice over IP (VoIP) protocols.

Supporting real-time applications over the Internet has many challenges [11]. Services such as VoIP require minimum service guarantees that go beyond the best-effort structure of today's IP networks. Though some codecs are capable of some level of adaptation and error concealment, VoIP quality remains sensitive to performance degradation in the network. Sustaining good quality VoIP calls becomes even more challenging when the IP network is extended to the wireless domain – either through 802.11 based wireless LANs or third generation (3G) cellular networks [5], [18], [25]. Such wireless extension of services are becoming more essential as there is already a huge demand for real-time services over wireless networks. Though, bare basic versions of services such as real-time news, streaming audio, and video on demand are currently being supported, the widespread use and bandwidth demands of these multimedia applications are far exceeding the capacity of current third generation (3G) cellular and wireless LAN technologies. Moreover, most access technologies do not have the option to differentiate specific application demands or user needs. With the rapid growth of wireless technologies, the task of providing broadband *last mile* connectivity is still a challenge. The last mile is generally referred to as a connection from a service provider's network to the end user – either a residential home or a business facility. Among the new wireless broadband access technologies that are being considered, WiMax (worldwide interoperability of microwave access) is perhaps the strongest contender that is being supported and developed by a consortium of companies [27].

A. WiMax

WiMax is a wireless metropolitan access network (MAN) technology that is based on the standards defined in the IEEE 802.16 specification. This standard-based approach is not only simplifying but also unifying development and deployment of WiMax. The 802.16 standard can be used in a point-to-point or mesh topology using pairs of directional antennas. These antennas can be used to increase the effective range of the system relative to what can be achieved in point-to-multipoint mode.

WiMax is envisioned as a solution to the outdoor broadband wireless access that is capable of delivering high-speed streaming data. It has the capability to deliver high-speed services upto a range of 30 miles, thus posing a strong competition to the existing last mile broadband access technologies such as cable and DSL. WiMax uses multiple channels for a single transmission and provides bandwidth of upto 100 Mbps [23]. The use of orthogonal frequency division multiplexing (OFDM) increases the bandwidth and data capacity by spacing channels very close to each other and still avoids interference because of orthogonal channels. A typical WiMax base station provides enough bandwidth to cater to the demands of more than fifty businesses with T1 (1.544 Mbps) level services and hundreds of homes with high-speed Internet access. WiMax's low cost of deployment coupled with existing demands from under-served areas creates major business opportunities.

B. Contributions of this paper

In this paper, we explore the possibility of supporting VoIP streams over WiMax and suggest means through which the quality of multiple VoIP streams can be improved. Specifically, the contributions of this paper are as follows.

- We show how the quality of VoIP calls is represented by R-score that primarily depends on loss and delay of VoIP packets. We show that loss is more sensitive than delay and hence try to recover as many dropped packets as possible at the cost of increased delay as long as the delay is within acceptable limits.

• We exploit the flexible features of the MAC layer of WiMax to dynamically construct and transmit the *MAC protocol data units* (MPDU) for supporting multiple VoIP streams over a single WiMax link.

• We use *aggregation* to construct variable-sized MPDUs based on the wireless channel conditions. We design a feedback mechanism at the MAC layer of the receiver which lets the transmitter know about the channel conditions. Depending on the feedbacks, the MAC layer at the transmitting side modifies its MPDU payload size and/or forward error correction code.

• The dynamic manner in which the MPDUs are changed to match the channel conditions and/or *minislots*, helps in increasing the packet restore probability, thereby increasing number of VoIP streams and their quality. The reduction in the number of retransmissions of dropped or corrupted packets lowers the delay, which is crucial for VoIP.

• We conduct simulation experiments to verify our proposed scheme. We assume a three-state Markovian channel model and study the performance with and without retransmissions. We show that the feedback-based technique coupled with retransmissions, aggregation, and variable length MPDUs are effective and increases the R-score by about 40%.

C. Organization

The rest of the paper is organized as follows. We provide a brief overview on the adaptive techniques that have been proposed to support data/streaming services over wireless channels in section II. In section III, we discuss the rich set of MAC layer features of WiMax with particular emphasis on aggregation and fragmentation. In section IV, we show the effect of delay and loss on R-score – a metric used to represent the quality of VoIP. We demonstrate that VoIP calls are more sensitive to loss than delay. Based on this observation, we propose our adaptive MPDU construction scheme in section V. In section VI, we present the simulation model and results. Conclusions are drawn in the last section.

II. Related Work

Supporting real-time applications over any wireless network (e.g., 3G cellular networks, IEEE 802.11 based wireless local area networks and IEEE 802.16 based WiMax) poses many challenges, including limited bandwidth, coping with bandwidth fluctuations, and lost or corrupted data. Due to the growing popularity of streaming services over wireless networks, the problems have been well researched and many solutions have been proposed that combine audio and video processing techniques with mechanism that are usually dealt with in the data link and physical layer. These approaches can be broadly classified into two categories – ARQ (automatic repeat request) and FEC (forward error correction). ARQ schemes provide high reliability when the channel is good or moderate. However, for error prone channels, the throughput drops due to increased frequency of retransmissions. In order to counter this effect, hybrid ARQ schemes are used that combine FEC with ARQ schemes.

As far as VoIP is concerned, an assessment of the Internet in supporting toll quality telephone calls was conducted in [16]. The assessment was based on delay and loss measurements that were taken over wide-area backbone networks, considering realistic VoIP scenarios. The findings indicate that although voice services can be adequately provided by some providers, a significant number of paths lead to poor performance even for excellent VoIP end-systems. The tuning of a codec for a particular type of network is very important. For example, an Adaptive Multi-Rate (AMR) voice codec was properly tuned for IEEE 802.16 networks that allowed switching to the maximum encoding rate [22]. Note, such codecs can also be tuned for other networks as well. The study in [7] presents a simulation model and analyzes the performance of a IEEE 802.16 system by focusing on the MAC layer scheduling for VoIP traffic using AMR codecs. However, for IP networks, the aggregate background traffic affects the performance of VoIP. In [6], a study was conducted where active and passive traffic measurements were taken to identify the issues involved with the deployment of voice services over the IP network. The results show that no QoS differentiation is needed in

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

the current backbone but new protocols and mechanisms need to be introduced to provide better protection against link failures. The reason is that link failures are followed by long periods of routing instability, during which packets are dropped because of being forwarded along invalid paths. The effect of bursty packet losses in the Internet was taken care of by changing the packet interval in [12]. Two loss repair methods – FEC and low bit-rate redundancy were used to improve the VoIP perceived quality. Through mean opinion score (MOS) test results, it was found that FEC performed better than bit-rate redundancy. In [4], localized packet loss recovery and rapid rerouting was used in the event of network failures for VoIP packets. The recovery protocols were deployed on the nodes of an application-level overlay network and required no changes to the underlying infrastructure. Retransmission strategies for VoIP packets were deployed using unsolicited grant service (UGS) scheduling in [17].

In this paper, we do not propose a new link layer technique. In-stead, we use the commonly used FEC and ARQ schemes and apply that to the MAC layer of WiMax. These techniques are so used that they do not contradict the MAC layer specifications that have already been defined for WiMax. The novelty of our approach lies in the exploitation of the features of both VoIP and WiMax for improving the quality of VoIP calls over WiMax channels.

III. THE MAC LAYER OF WIMAX

WiMax offers some flexible features that can potentially be exploited for delivering real-time services. In particular, though the MAC layer of WiMax has been standardized, there are certain features that can be tuned and made application and/or channel specific [2], [21]. For example, the MAC layer does not restrict itself to fixed-size frames, but allows variable-sized frames to be constructed and transmitted. Let us first discuss the MAC layer of WiMax.

The MAC layer of WiMax comprises three sub-layers which interact with each other through the service access points (SAPs) as shown in figure 1. The service specific convergence sub-layer provides the transformation or mapping of external network data, with the help of the service access

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

point. The MAC common part sublayer receives this information in the form of MAC service data units (MSDUs) which are packed into the payload fields to form MAC protocol data units (MPDUs). Privacy sublayer provides authentication, secure key exchange and encryption on the MPDUs and passes them over to the physical layer. Of the three sublayers, the common part sublayer is the core functional layer which provides bandwidth, and establishes and maintains connection. Moreover, as the WiMax MAC provides a connection-oriented service to the subscriber stations, the common part sublayer also provides a connection identifier to identify which connection the MPDU is servicing.

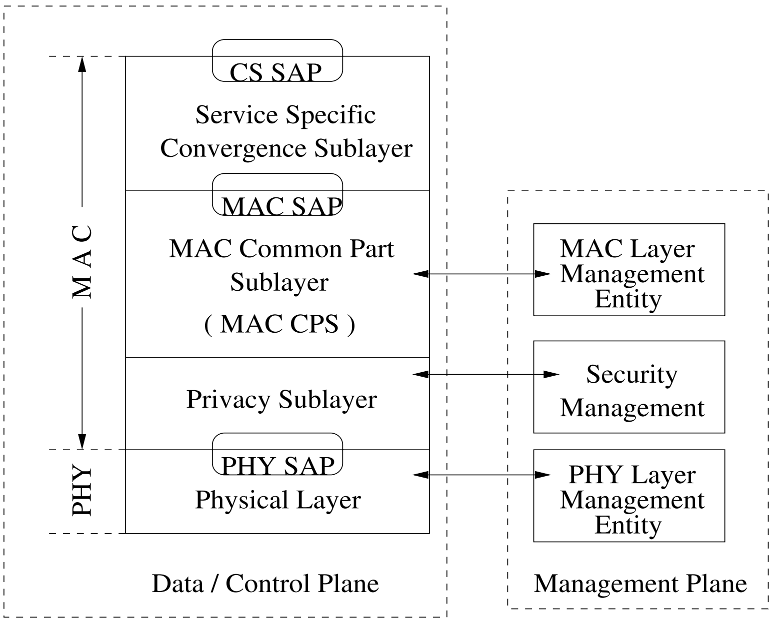


Fig. 1. WiMax MAC layer with SAPs

Let us discuss the common part sublayer and its rich set of features. This sublayer controls the on-air timing based on consecutive frames that are divided into time slots. The size of these frames and the size of the individual slots within these frames can be varied on a frame-by-frame basis. This allows effective allocation of on-air resources which can be applied to the MPDUs to be transmitted. Depending on the feedback received from the receiver and on-air physical layer slots, the size of the MPDU can be optimized. In the research, we exploit this feature of the common part sublayer that modifies the size of the MPDUs to adapt to the varying channel conditions.

A. Aggregation

The common part sublayer is capable of packing more than one complete or partial MSDUs into one MPDU. In figure 2, we show how the payload of the MPDU can accommodate more than two complete MSDUs, but not three. Therefore, a part of the third MSDU is packed with the previous two MSDUs to fill up the remaining payload field preventing wastage of resources. The payload size is determined by on-air timing slots and feedback received from subscriber station.

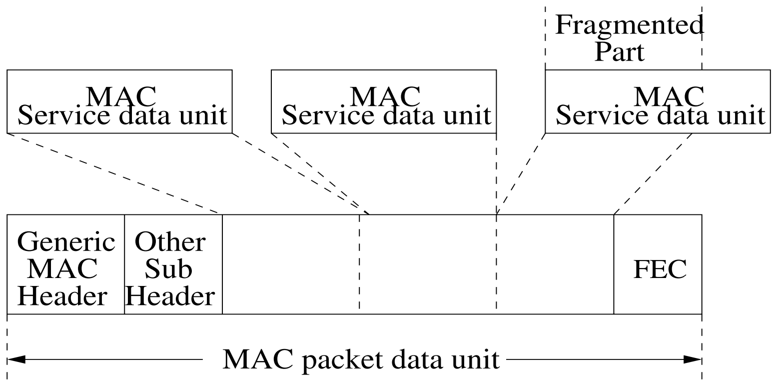


Fig. 2. An MPDU accommodating multiple MSDUs

B. Fragmentation

The common part sublayer can also fragment a MSDU into multiple MPDUs. In figure 3, we show how a portion of a single MSDU occupies the entire payload field of a MPDU. Here, the payload field of the MAC packet data unit to be transmitted is too small to accommodate a complete MSDU. In that case, we fragment a single MSDU and pack the fragmented part into the payload field of the MPDU.

IV. Delay and Loss Sensitivity of VoIP

As VoIP packets travel through a network, there are evidently some congestion and channel related losses. Also, the packets suffer delay depending on the congestion at the intermediate routers. Both

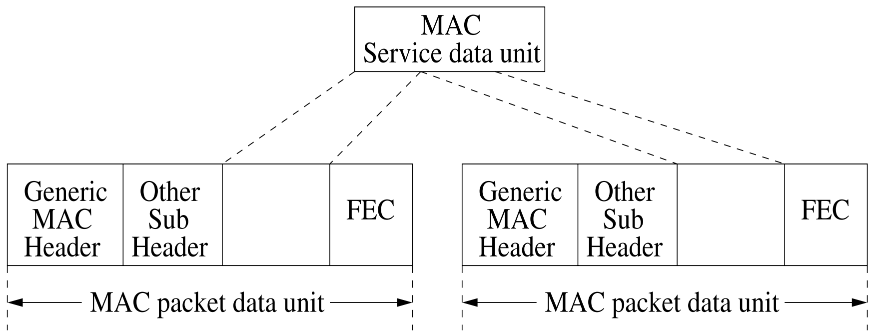


Fig. 3. Single MSDU fragmented to multiple MPDUs

loss and delay of packets adversely affect the quality of VoIP calls which is generally expressed in terms of R-score.

A. Quality of VoIP and R-Score

A typical VoIP application works as follows. First, a voice signal is sampled, digitized, and encoded using a given algorithm/coder. The encoded data (called frames) is packetized and transmitted using RTP/UDP/IP. At the receiver's side, data is de-packetized and forwarded to a playout buffer, which smoothes out the delay incurred in the network. Finally, the data is decoded and the voice signal is reconstructed.

The quality of the reconstructed voice signal is subjective and therefore is measured by the mean opinion score (MOS). MOS is a subjective quality score that ranges from 1 (worst) to 5 (best) and is obtained by conducting subjective surveys. Though these methods provide a good assessment technique, they fail to provide an on-line assessment which might be made use of for adaptation purpose. The ITU-T E-Model [3] provides a parametric estimation and defines an *R-factor* that combines different aspects of voice quality impairment. It is given by

$$R = 100 - I_s - I_e - I_d + A \tag{1}$$

where I_s is the signal-to-noise impairments associated with typical switched circuit networks paths, I_e is an equipment impairment factor associated with the losses due to the codecs and network,

I_d represents the impairment caused by the mouth-to-ear delay, and A compensates for the above impairments under various user conditions and is known as the expectation factor.

We note that the contributions to the R-score due to delay and loss impairments are separable. This does not mean that the delay and loss impairments are totally uncorrelated, but only their influence can be measured in isolated manner. Expectation factor covers intangible and almost impossible to measure quantities like expectation of qualities. However, no such agreement on measurement of expectation on qualities has still been made and for this reason expectation factor is usually dropped from the R-factor in most studies. The R-factor ranges from 0 to 100 and a score of more than 70 usually means a VoIP stream of decent quality. The R-score is related to MOS through the following non-linear mapping [3]

$$MOS = 1 + 0.035R + 7 \times 10^{-6} R(R - 60)(100 - R) \tag{2}$$

for $0 \leq R \leq 100$. If $R < 0$, MOS takes the value of 1, and similarly, if $R > 100$, MOS takes the value of 4.5.

Among all the factors in equation (1), only I_d and I_e are typically considered variables in VoIP [8]. Using default values for all other factors, the expression for R-factor given by equation (1) can be reduced to [3]

$$R = 94.2 - I_e - I_d \tag{3}$$

B. Delay and Loss Sensitivity of VoIP

Let us study how end-to-end delay (consisting of codec delay, network delay, and playout delay) and loss probability (consisting of loss in the network and playout loss at the receiver's decoder buffer) affect the VoIP call quality, i.e., the R-score.

1) Effect of Delay: In a VoIP system, the total mouth-to-ear delay is composed of three components: codec delay (d_{codec}), playout delay ($d_{playout}$), and network delay ($d_{network}$). Codec delay represents the algorithmic and packetization delay associated with the codec and varies from codec to codec. For example, the G.729a codec introduces a delay of 25 ms. Playout delay is the delay associated with the receiver side buffer required to smooth out the jitter for the arriving packet streams. Network delay is the one-way transit delay across the IP transport network from one gateway to another. Thus the total delay is

$$d = d_{codec} + d_{playout} + d_{network} \tag{4}$$

The delay impairment, denoted by I_d, depends on the one way mouth-to-ear delay experienced by the VoIP streams. This mouth-to-ear delay determines the interactivity of voice communication. Its impact on voice quality depends on a critical time value of 177.3 ms [8], which is the total delay budget (one way mouth-to-ear delay) for VoIP streams. The effect of this delay is modeled as

$$I_d = 0.024d + 0.11(d - 177.3)\mathbf{H}(d - 177.3) \tag{5}$$

where $\mathbf{H}(x)$ is an indicator function: $\mathbf{H}(x) = 0$ if $x < 0$, and 1 otherwise.

2) Effect of Loss: VoIP call quality is also dependent on the loss impairment. Recall, I_e represents the effect of packet loss rate. I_e accounts for impairments caused by both network and receiver's playout losses. Different codecs with their unique encoding/decoding algorithms and packet loss concealment techniques yield different values for I_e. We use the E-model as proposed in [3], [8], [9] that relate I_e to the overall packet loss rate as

$$I_e = \gamma_1 + \gamma_2 ln(1 + \gamma_3 e) \tag{6}$$

where γ_1 is a constant that determines voice quality impairment caused by encoding, and γ_2 and γ_3 describe the impact of loss on perceived voice quality for a given codec. Note that e includes both network losses and playout buffer losses, which can be modeled as,

$$e = e_{network} + (1 - e_{network})e_{playout} \tag{7}$$

where, $e_{network}$ is the loss probability due to the loss in the network and $e_{playout}$ is loss probability due to the playout loss at the receiver side.

The values of the well known parameters $(\gamma_1, \gamma_2, \gamma_3)$ for G.729a and G.711 codecs are shown in table I [3].

Codec	γ_1	γ_2	γ_3
G.729a	11	40	10
G.711	0	30	15

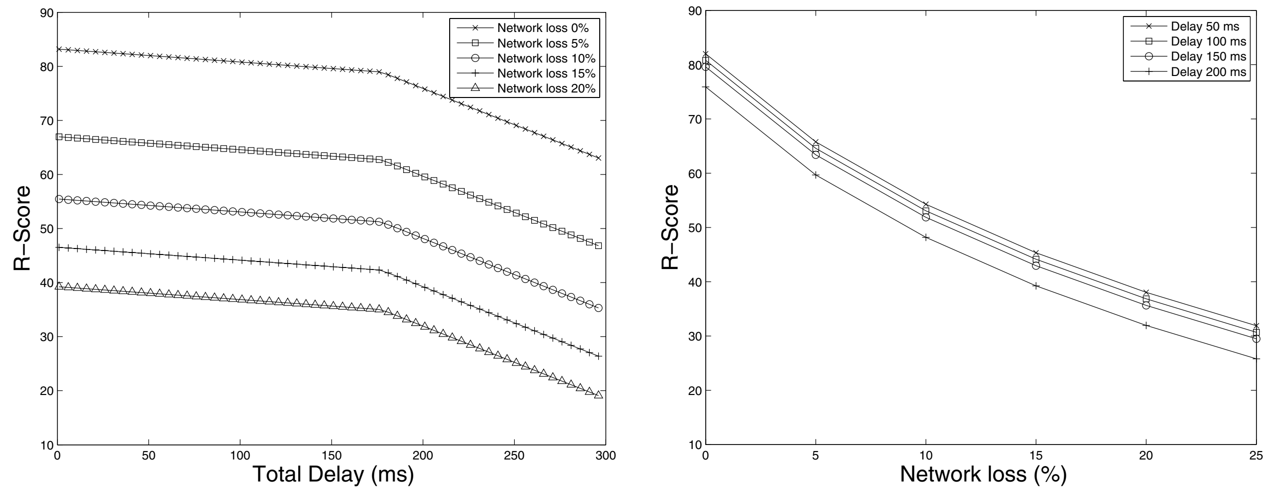
TABLE I

LOSS IMPAIRMENT PARAMETERS [3]

3) Sensitivity of R-score towards delay and loss: We rewrite equation (3) using equation (5) and (6) as

$$R = 94.2 - (\gamma_1 + \gamma_2 ln(1 + \gamma_3 e)) - (0.024d + 0.11(d - 177.3)\mathbf{H}(d - 177.3)) \tag{8}$$

To find the sensitivity of loss and delay towards R-score of VoIP calls, we use different delay values keeping the network loss fixed in equation (8). In figure 4(a), we observe how the R-score changes with increasing delay for fixed loss. Similarly, in figure 4(b), we show how the R-score changes with loss for fixed delay. From figure 4(a) we observe that there is no significant drop in R-score for a



(a) (b)

Fig. 4. a) R-score vs. delay; b) R-score vs. network loss

given network loss rate. R-score drops relatively rapidly when the delay exceeds 177.3 ms. Note, this is the delay threshold defined in [3]. However, when loss in increased from 0% to 5%, the drop in R-score is about 14. Also, from figure 4(b), we see that the drop in R-score is mainly due to loss (x-axis). When delay is significantly increased from 50 ms to 200 ms, the drop in R-score is about 5. Thus we can infer that loss is more crucial than delay. This difference in sensitivity motivates us to manipulate the loss and delay. Next, we propose an adaptive mechanism that is implemented at the MAC layer of WiMax. Our objective is to recover as many dropped packets as possible to minimize the loss probability at the cost of increased delay as long as the delay is within acceptable limit.

V. Adaptive MPDU Construction

With the sensitivity of VoIP with respect to loss and delay known, we devise adaptive schemes at the MAC layer to dynamically construct the MPDUs. Once a connection is set-up, the size of every MPDU is determined such that it strikes a balance between the lost packets and the delay incurred. Our aim is to improve the quality of VoIP calls and at the same time increase the number of streams that can be accommodated.

A. Connection Set-up and Transmission

Let us now discuss how a new connection is set-up and how the MPDUs are transmitted. We consider a WiMax base station providing services, including VoIP calls, to the subscriber stations in that cell. For this research, we only consider the VoIP calls and assume that the base station handles multiple VoIP calls simultaneously. We ignore the source routers for each call, but the destinations are all within the cell. Effectively, the last hop of the VoIP path is the WiMax link that provides the wireless coverage. The identity of each call is maintained by the connection identifier (CID) provided by the common part sublayer. As a result, VoIP packets (which are inherently very small) do not have to deal with contention overhead, which greatly increases the efficiency, i.e., number of VoIP streams.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

Phase 1: *Subscriber station requests connection request:*

Subscriber station that wants a VoIP service stream from the base station transmits the ranging request (RNG-REQ) packet that enables the base station to identify the initial ranging, timing and power parameters. Service flow parameters requests (bandwidth, frequency, peak or average rate) are sent next and variable length MSDU indicators are turned on.

Phase 2: *Base station confirms connection:*

After receiving connection request from a subscriber station, the base station transmits a ranging response which provides the initial ranging, timing and power adjustment information to the subscriber station. VoIP service flow parameters are agreed on and a basic connection-ID is provided to the subscriber station.

Phase 3: *Base station starts transmission of MPDUs:*

MSDUs obtained from the MAC convergence sublayer are converted to MPDUs. As needed, MSDUs can either be packed or fragmented to form the desired sized MPDUs. Since no feedback is received at the start of transmission, the payload and code size agreed at the time of connection establishment is maintained. When a feedback is received, the next awaiting MPDU is formed depending on the type of feedback received. On the reception of the feedbacks, the payload and code sizes are changed. We note that the increase or decrease in payload and code will depend on the ratio of the payload and code.

B. Packet Restore Probability

If a receiver gets a corrupted packet, it is in no position to correct the errors. However, if some redundant bits in the form of forward error correction are applied before transmission, then there is a probability that the receiver would be able to detect and correct the errors. The correction capability of these codes will depend on the kind and the length of the codes used. Since this paper does not attempt to propose new coding techniques, we will just use the simplest of codes – *block codes*.

In block codes, M redundancy bits are added to N information bearing bits. (Note that these extra bits are generated using a generator matrix operating on the bits.) If we consider such a MPDU, the resulting bit loss probability is given by [26],

$$b = \sum_{i=M+1}^{M+N} \binom{M+N}{i} {b_p}^i (1-b_p)^{M+N-i} \frac{i}{M+N} \tag{9}$$

where, b_p is the bit loss probability before decoding and b is the decoded bit error probability. The restore probability of such a MPDU with payload size N bits and code size M bits is given by, $p = (1-b)^{(M+N)}$. We apply three schemes to manipulate this packet restore probability:

1) Decrease payload with code size fixed: Let b be the resulting bit loss probability after decoding of a MPDU with payload size N and code size M. Now, if we decrease the payload size to N' ($N' < N$) keeping the code size fixed, then the resulting bit loss probability after decoding is given by

$$b' = \sum_{i=M+1}^{M+N'} \binom{M+N'}{i} {b_p}^i (1-b_p)^{M+N'-i} \frac{i}{M+N'} \tag{10}$$

As evident from coding theory, with decrease in payload size with fixed redundancy codes, the decoded bit error probability decreases resulting in $b' < b$. As far as packet restore probability is concerned, with modified payload size N', p' is given by,

$$p' = (1-b')^{(M+N')} \tag{11}$$

The ratio of p' to p is given by

$$\begin{aligned} \frac{p'}{p} &= \frac{(1-b')^{(M+N')}}{(1-b)^{(M+N)}} \\ &= \Big(\frac{1-b'}{1-b}\Big)^{(M+N)} \times \frac{1}{(1-b')^{(N-N')}} \end{aligned} \tag{12}$$

As $b' < b$, $(1-b') > (1-b)$. Thus the first term in the equation (12), $\left(\frac{1-b'}{1-b}\right)^{(M+N)} > 1$. Again, as $(1-b') < 1$, the second term, $\frac{1}{(1-b')^{(N-N')}}$ is also greater than 1. This establishes that $\frac{p'}{p} > 1$, resulting in $p' > p$.

2) Increase code size with payload fixed: Similarly, it can be argued that if the code is increased keeping the payload fixed, the resulting bit loss probability decreases and packet restore probability of MPDUs increases.

3) Increase both payload and code size: The third scheme would be to increase both the payload and the code size. As we know, increasing payload only will increase the resulting bit error probability, so we must also increase the code to compensate for the increased payload.

C. Enabling ARQ mechanism

Though the application of FEC enhances packet restore probability, the performance can still be further improved if the optional Automatic Repeat Request (ARQ) mechanism is enabled. The ARQ mechanism at the WiMax MAC common part sublayer is enabled by the exchange of control messages between the transmitter and the receiver at the time of connection set up. The ARQ allows feedback to be received at the transmitter side to understand the ongoing call quality and the channel status. We enable the ARQ mechanism and make every subscriber station send a feedback in terms of the packet restore probability, from which WiMax MAC common part sublayer gets the information if a packet has been received successfully or not. In addition, these feedbacks give an estimate about the channel status.

We apply fast feedback at the MAC layer and use very small packets to reduce overhead. The parameters used in the feedback packets are connection identifier (CID), ARQ status (enabled or disabled), maximum retransmission limit, packet restore probability, and a sequence number. The sequence number is used to correlate packets with its response from the base station. If the packet is not received correctly, i.e., the packet restore probability is below a certain threshold, then retransmission mechanism is applied. The main advantage of using the retransmission scheme is to lower the loss impairment at the expense of increased delay. For MAC layer retransmissions, we maintain a buffer for every stream at the transmitting WiMax base station. This buffer helps in temporarily storing the packets unless and until the packets are restored correctly by the receiver. This of course introduces a delay which we denote by d_{queue}. Thus the total one-way mouth-to-ear delay, as previously given

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

by equation (4), is modified as

$$d = d_{codec} + d_{playout} + d_{network} + d_{queue} \tag{13}$$

To counter this increase in delay, we use *aggregation*. We make use of this feature to pack multiple MSDUs into one MPDU thus making optimal MPDU size.

D. Optimal MPDU Size

Since packets often get lost or corrupted during transmission in error prone wireless channels, ARQ mechanism is usually used to identify and possibly recover the missing frames. In our case, ARQ will play a crucial role in estimating the channel condition and the fate of the MPDUs that have been transmitted. As a result, the round trip time (RTT) becomes crucial in determining the size of the MPDUs. We define RTT as the time difference between the time the last bit of a MPDU is transmitted and the time the acknowledgment for that MPDU is received. Moreover, we assume zero time interval between the transmissions of two consecutive MPDUs, i.e., the last bit of a MPDU and the first bit of the next MPDU are transmitted back to back.

Let us now show, how the RTT affects the size of the MPDUs. If we assume that t is the time taken to transmit the MPDU and T is the RTT, then the number of MPDUs already transmitted before the acknowledgment of the first MPDU received is given by $\lceil T/t \rceil$. It can be noted that t depends on the size of the MPDU and thus there is a trade-off between the goodput (information bits/total bits transmitted) and the delay. If a MPDU is large, the transmission time is large but the overhead due to headers is less which helps in maintaining a high goodput. If a MPDU is dropped or corrupted due to bad channel condition, the ARQ mechanism will trigger the retransmission of the large MPDU, which will increase the delay in the transmission. Moreover, by the time the MAC common part sublayer receives the feedback i.e., learns about the channel condition, the transmission of the next MPDU would have already started. If the bad channel condition persists, the probability of the subsequent frame being dropped or corrupted is very high. Thus, there will be more retransmissions of large

MPDUs under bad channel condition, resulting in severe degradation of goodput compromising the QoS. On the other hand, if the MPDU size is too small, the transmission time will be less but the main disadvantage of having small MPDUs is the low goodput due to low payload/overhead ratio. Thus, we observe that both large and small MPDUs have their advantages and disadvantages. We propose to combine the advantages of both by dynamically changing the MPDU size in response to the type of the feedback and allocation of minislots for VoIP streams to obtain a desired level of performance. The pseudo-code for the adaptive MPDU construction is presented in algorithm 1 (n_1, m_1, m_2, n_3, m_3 are implementation dependent parameters).

Feedback type	MPDU status at the receiver
1	MPDU received correctly
2	MPDU received with errors, and uncorrectable
3	MPDU dropped, timeout in receiver MAC occurred
4	Receiver MAC buffer full

TABLE II

FEEDBACK CLASSIFICATION

E. Dynamic Allocation of Minislots

The users in a WiMax cell are serviced in a TDMA/TDD manner after their connections are set up. One or multiple minislots are assigned to every user to service their requests. More formally a minislot is defined as a unit of uplink/downlink bandwidth allocation equivalent to n physical symbols, where $n = 2^m$ and m is an integer ranging between 0 and 7. The number of physical symbols within each frame is a function of the symbol rate. The symbol rate is selected in order to obtain an integral number of physical symbols within each frame. For example, with a 20 Mbaud symbol rate, there are 5000 physical symbols within a 1 ms frame.

In addition to the already proposed mechanisms, we propose another key aspect of dynamic minislot allocation for not only enhancing the performance of the VoIP calls but also accommodating more

Algorithm 1 Feedback based adaptive MPDU construction

```
Input: Feedback classification Table II
For each transmitted MPDU i {
    BS receives a feedback from the receiver

    IF (feedback type == 1) {
        MPDU i is flushed from BS buffer
        Aggregate leading MSDUs in queue to increase payload by n_1 bytes
        Decrease FEC size by m_1 bytes
    }

    IF (feedback type == 2) {
        IF (retransmit count ≤ MAX_Retransmission_Count) {
            Retransmit MPDU i
            Keep payload size fixed
            Increase FEC size by m_2 bytes
        }
        ELSE {
            Discard MPDU i
        }
    }

    IF (feedback type == 3) {
        IF (retransmit count ≤ MAX_Retransmission_Count) {
            Aggregate leading MSDUs in queue to decrease payload by n_3 bytes
            Increase FEC size by m_3 bytes
            Retransmit MPDU i
        }
        ELSE {
            Discard MPDU i
        }
    }

    IF (feedback type == 4) {
        Stall transmission for a certain period
    }
}
```

number of calls. For a G.729a codec, a typical VoIP packet is 60 bytes (40 bytes RTP/UDP/IP header and 20 bytes payload) which is fed to the WiMax MAC layer. At the MAC layer, a minimum overhead is introduced (generic MAC header of 6 bytes for data MPDUs) and FEC codes (depending on number of retransmissions of this frame and codec efficiency) are appended for error recovery. Thus transmission of a MPDU (consisting of a single MSDU) takes about $8-10$ μs. On the other hand, minimum and maximum minislot duration are 1 physical symbol (0.2 μs) and 128 physical

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

symbols (≈ 26 μs) respectively with 20 Mbaud symbol rate. Thus the duration of minislot allocated plays a vital role for VoIP packets. If a single minislot of duration less than the minimum MPDU size is allocated to a session then there is no way that the MPDU can be accommodated in that minislot. Hence this kind of single slot allocation cannot be put to effective use. The better option is to allocate multiple minislots to a single user to avoid wastage of minislots. Now the question arises how many minislots should be assigned to a single user and what scheduling policy should be used to reduce the delay impairment. As each VoIP stream has a delay budget (177.3 ms), the scheduling policy must consider the delay that a VoIP stream has already suffered. Therefore, we use a scheduling policy where the base station looks up its buffers for respective streams and calculates the delay of the leading MPDUs in each stream and assigns the minislots to the stream that has suffered the highest delay. The number of minislots assigned are such that the duration of all the combined minislots are greater than or equal to the MPDU(s) being transmitted.

VI. Simulation Model and Results

We conducted simulation experiments to evaluate the improvements achieved by the proposed techniques. Evaluations for adaptive and non-adaptive schemes were done under the same channel conditions for a fair comparison.

A. Channel Model

We assumed a three-state Markov model for the channel. Three states were used to have more granularities in the channel conditions. Each state was characterized by a certain bit error probability (BER): *good state* had a BER of 0.01, the *medium state* had a BER of 0.07, and the *bad state* had a BER of 1.0. By setting appropriate transition probabilities among these three states, we are able to model different channel conditions for our simulation.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

B. Simulation Parameters for VoIP

For our simulation, we assumed that the VoIP streams were generated by a G.729a codec. Note that, we could have used the specifications of any other voice codec (e.g., G.711 or AMR). However, our intention is to demonstrate how the MAC features of WiMax can be exploited for VoIP streams irrespective of their encoding techniques. Other simulation parameters are shown in the Table III.

Simulation Parameters	Values
d_{codec}	25 ms
$d_{playout}$	60 ms
$d_{network}$	70 ms
$e_{playout}$	0.005
WiMax minislot	2^m PHY symbols
m	0 - 7
1 ms WiMax frame	5000 PHY symbols
Symbol rate	20 Mbaud
WiMax bandwidth	100 Mbps

TABLE III

SIMULATION PARAMETERS

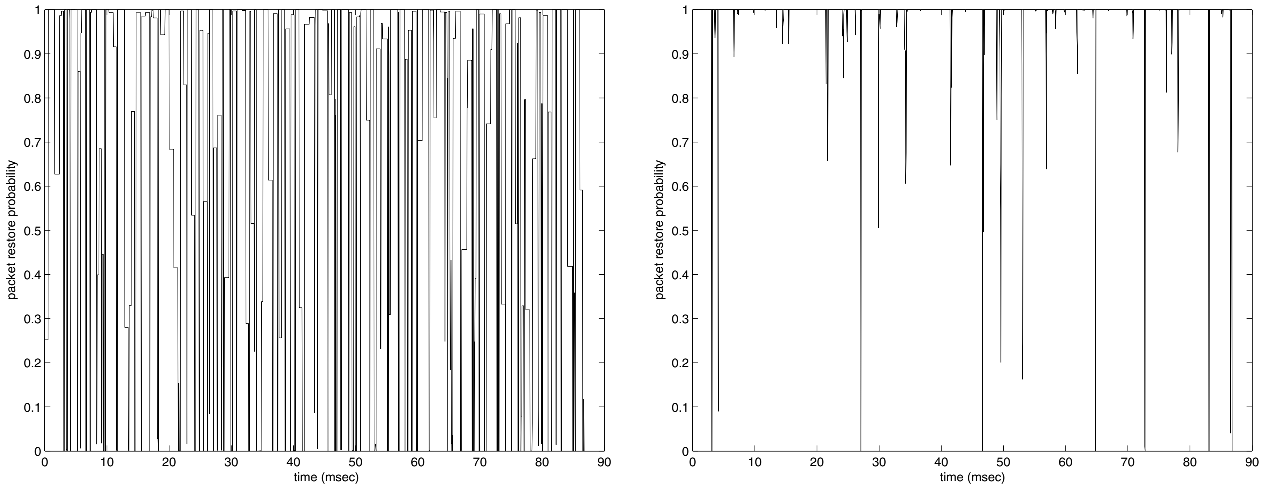
C. Simulation Results for VoIP

In figures 5(a) and 5(b), we present the packet restore probability for both non-adaptive and adaptive scheme. For this simulation we assumed that the channel remains in the good, medium, and bad state for 30%, 50% and 20% of the time respectively.

In the adaptive scheme, we used both the ARQ and aggregation schemes. In the non-adaptive scheme, we disabled the ARQ mechanism. With the adaptive scheme, packet restore probability is improved significantly.

In figure 6(a), we present the R-score for both adaptive and non-adaptive schemes without competing traffic. It is seen that with the adaptive scheme, there is an improvement of about 40% in

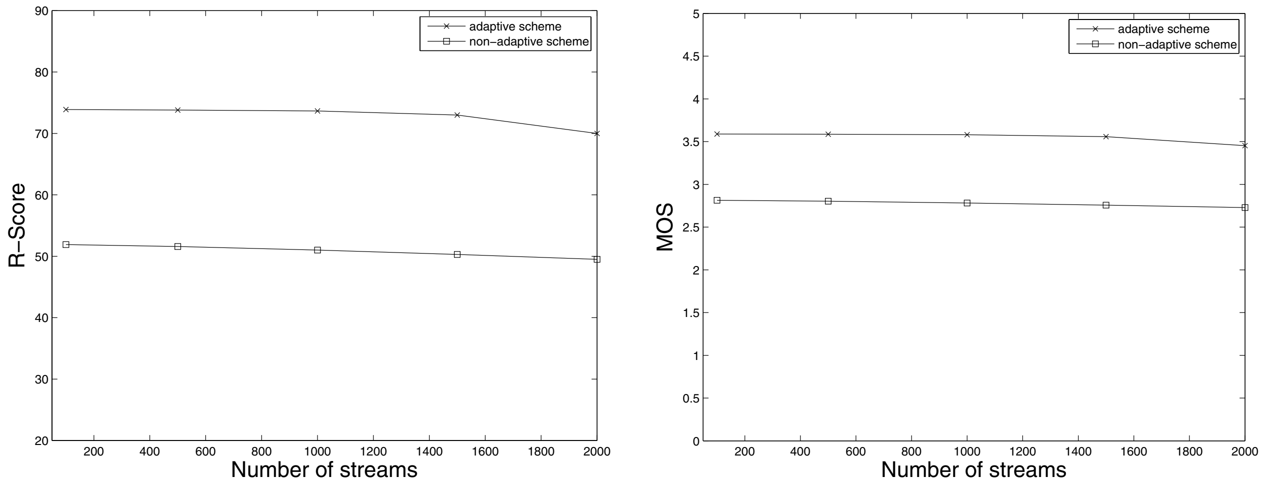
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.



(a) (b)

Fig. 5. a) Packet restore probability for non-adaptive scheme; b) Packet restore probability for adaptive scheme

R-score, which indicates that the call quality can be increased in WiMax using aggregation and ARQ on top of MAC common part sublayer. It can also be noted that with 2000 streams, R-score is still above 70, while with 1500 streams it is above 73.

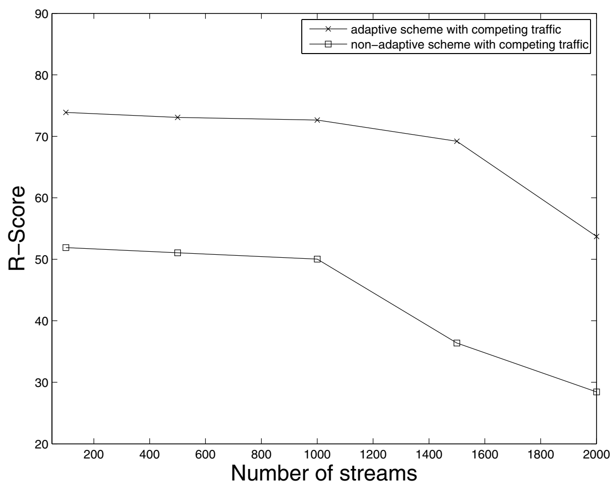


(a) (b)

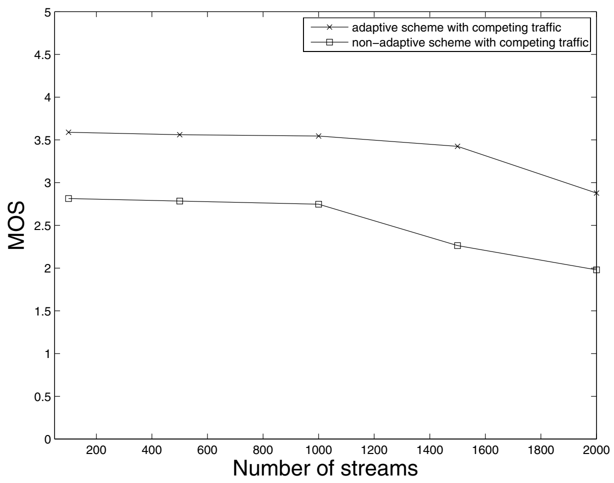
Fig. 6. a) R-score vs. number of VoIP streams; b) MOS vs. number of VoIP streams

In figure 6(b), we present the mean opinion score (MOS) for the same adaptive and non-adaptive schemes without competing traffic. It is observed that with adaptive scheme, MOS increases significantly (above 3.5) indicating the improvement of call quality of VoIP streams.

We have also modified our simulation model and introduced intermittent data sessions which generate competing traffic for the VoIP streams. The additional plots with competing traffic are shown in figures 7(a) and 7(b). We observe that even with intermittent data traffic, R-score and MOS for the adaptive scheme produces better result than with the non-adaptive scheme.



(a)

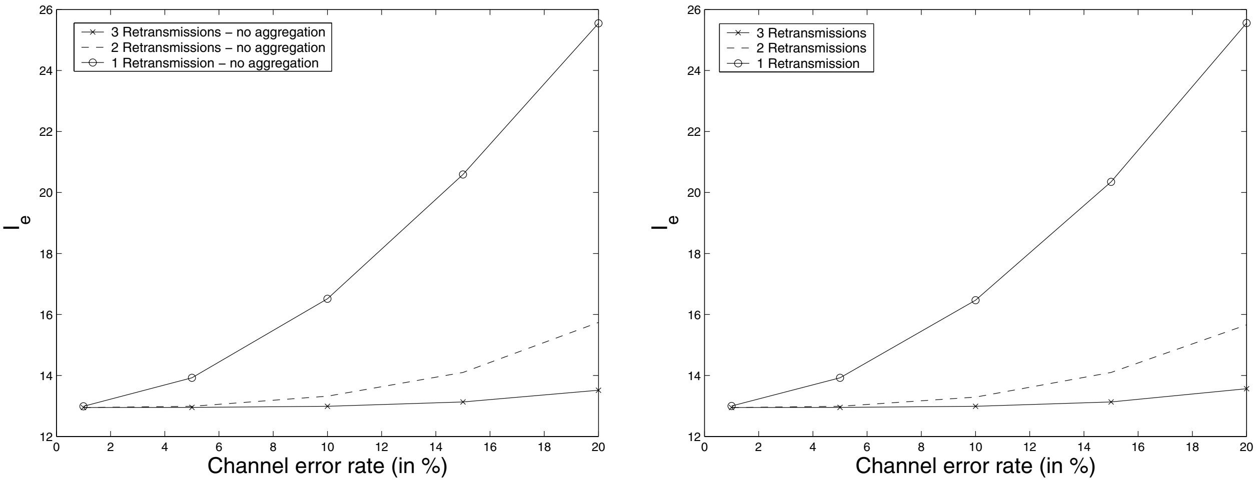


(b)

Fig. 7. a) R-score with competing data traffic; b) MOS with competing data traffic

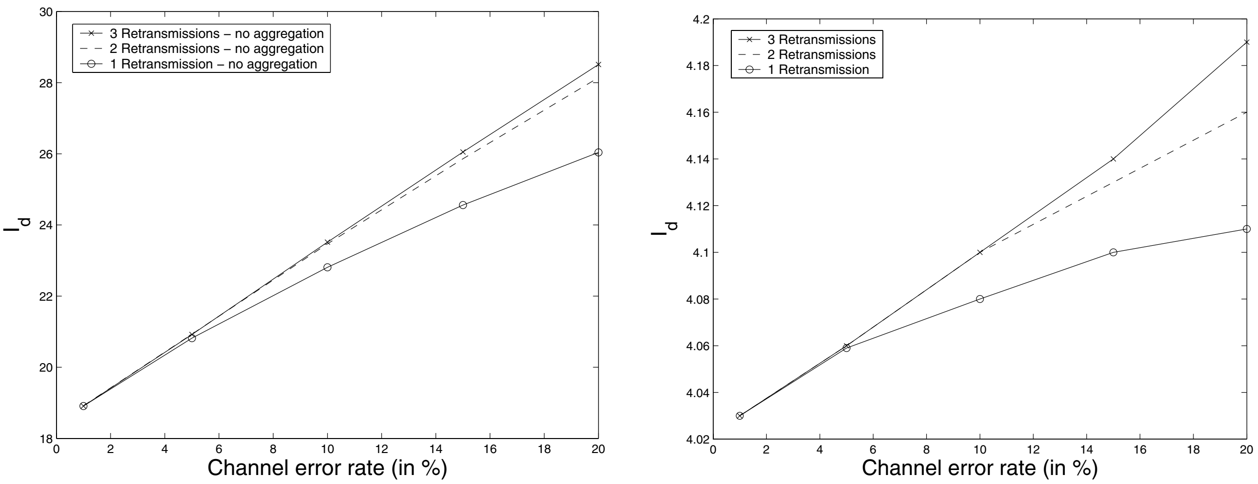
Next, we varied the maximum number of retransmissions from 1 to 3. It can be noted that even after the allowed number of retransmissions, a packet might not be restored. In that case, the packet is dropped. With such retransmission schemes, we study the loss and delay impairment. We fixed the number of VoIP streams at 1000 and gradually increased the channel error rate. Retransmission with aggregation and retransmission without aggregation are studied separately.

It is seen from figures 8(a) and 8(b) that loss impairment (I_e) is hardly affected by the introduction of aggregation. However, the delay impairment (I_d) is greatly reduced by the introduction of aggregation as seen in figures 9(a) and 9(b). (Note the difference in the range of Y-axes in figures 9(a) and 9(b)). From figures 8(b) and 9(b), we find that the loss impairment is more with the increase in channel error rate than the delay impairment for the retransmission with aggregation. On the other



(a) (b)

Fig. 8. a) Loss impairment without aggregation; b) Loss impairment with aggregation

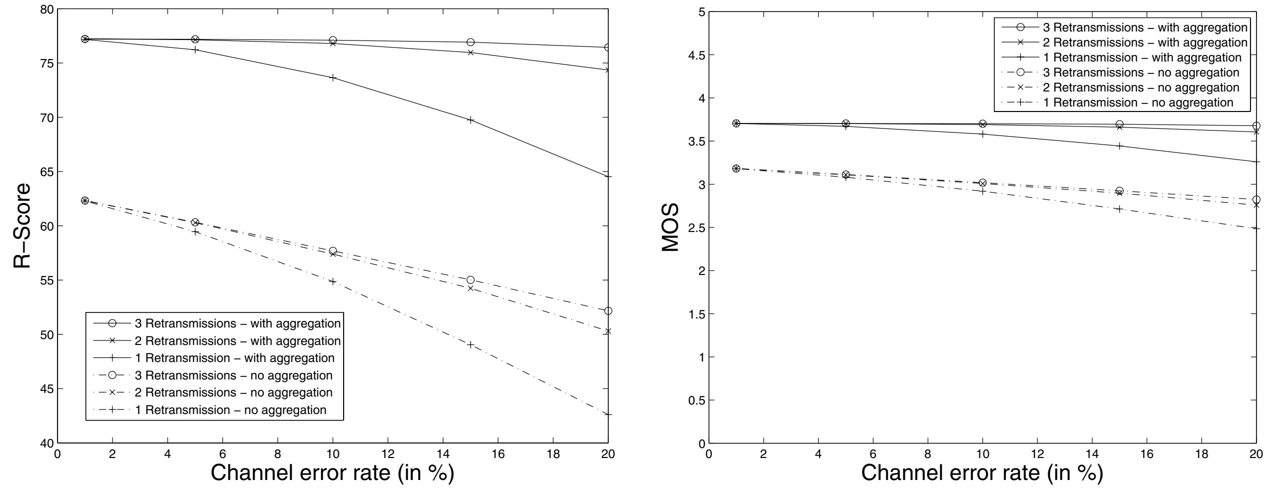


(a) (b)

Fig. 9. a) Delay impairment without aggregation; b) Delay impairment with aggregation

hand, for the retransmission without aggregation, delay impairment is more than loss impairment. This is because the requested minislot(s) are not fully utilized when packets are retransmitted without aggregation.

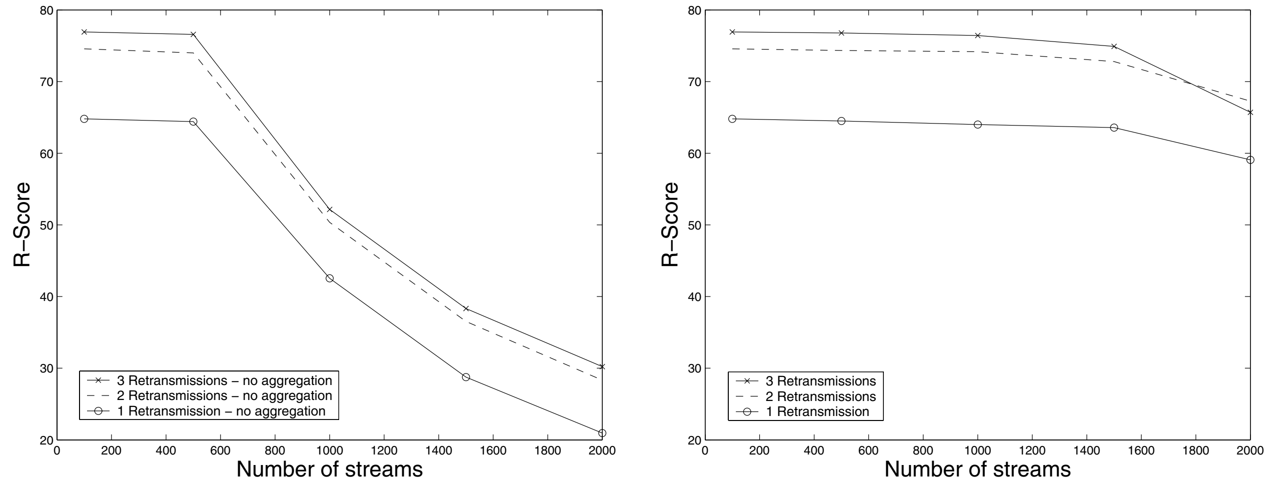
In figures 10(a) and 10(b), we present the variation of R-score and mean opinion score (MOS) vs. channel error rate with and without aggregation. It is evident from figures 10(a) and 10(b), that there is an improvement in R-score and MOS particularly when the allowed numbers of retransmissions are



(a) (b)

Fig. 10. a) R-score vs. error rate with and without aggregation; b) MOS vs. error rate with and without aggregation

2 and 3. It is observed that with the increase in channel error rates, the rate of decrease is much less for the 2 or 3 retransmissions than just 1 retransmission. It is also evident that the retransmission with aggregation scheme gives better R-score and MOS values than the retransmission without aggregation. Thus it is desirable to bundle both the features (retransmission and aggregation) in WiMax to improve the call quality in VoIP.



(a) (b)

Fig. 11. a) R-score vs. number of streams without aggregation; b) R-score vs. number of streams with aggregation

In figures 11(a) and 11(b), we show how the R-score is affected when the number of streams is increased. Channel error rate is assumed to be 20%. As expected, retransmission coupled with aggregation yields better R-score. In figure 11(a), R-score is presented without aggregation, while in figure 11(b), R-score is presented with aggregation. Aggregation helps putting multiple packets (MSDUs) in the queue ready to be transmitted; this introduces a lower delay impairment for the VoIP streams than without the aggregation feature enabled. Thus we show that with introducing the aggregation scheme, we can afford more number of VoIP streams without affecting their call quality.

Moreover, it is seen that with the retransmission with aggregation scheme, the 3-retransmission scheme gives better performance for low and medium load than 2 and 1-retransmission, but the performance degrades with increase in number of streams. The reason behind this is that in a typical cell, the VoIP streams share the common backhaul bandwidth. With more number of streams, the bandwidth allocation for each stream decreases. On the other hand, with increase in number of retransmissions and number of streams, there will be more packets in the backlog queue ready to be transmitted or retransmitted. This eventually increases the delay for the packets and introduces jitter at the receiving end; thus increasing delay impairment. As R-score decreases with increasing delay impairment, we find that 3-retransmission produces worse call quality with increasing load.

VII. Conclusions

As new wireless access technologies are being developed, WiMax is emerging as one of the promising broadband technologies that can support a variety of real-time services. Since extension of VoIP calls over wireless networks are inevitable, we study the feasibility of supporting VoIP over WiMax.

We propose a combination of techniques that can be adopted not only to enhance the performance of VoIP but also to support more number of VoIP calls. The proposed schemes, make use of the flexible MAC features – mainly the size of the protocol data units. We enable the ARQ, use FEC, construct MPDUs by aggregating multiple MSDUs, and dynamically allocate one or multiple minislots to every

VoIP call. The performance of the VoIP calls are studied with respect to R-score. We exploit the difference in sensitivity of R-score towards loss and delay for recovering as many packets as possible at the cost of increased delay. Exhaustive simulation experiments reveal that the feedback-based technique coupled with retransmissions, aggregation, and variable size MPDUs not only increases the the R-score (and consequently the MOS) but also the number of VoIP streams.

ACKNOWLEDGEMENTS

The authors are thankful to the anonymous reviewers for their useful comments in improving the quality and presentation of this paper.

REFERENCES

[1] IEEE Std. 802.16-2001 IEEE Standard for Local and MAN Part 16:Air Interface for Fixed Broadband Wireless Access Systems, 2002.

[2] IEEE Std. for Local and MAN Part 16:Air Interface for Fixed Broadband Wireless Access Systems- Amendment 2: MAC Modifications and Additional Physical Layer Spec. for 2-11 GHz, Std 802.16a-2003 (Amendment to IEEE Std 802.16-2001), 2003.

[3] ITU-T Recommendation G.107, “The E-Model, a computational model for use in transmission planning”, December 1998.

[4] Y. Amir, C. Danilov, S. Goose, D. Hedqvist, A. Terzis, “An Overlay Architecture for High-Quality VoIP Streams”, IEEE Transactions on Multimedia, vol. 8, Issue 6, Dec. 2006, pp. 1250-1262.

[5] W. Bang, K.I. Pedersen, T.E. Kolding, P.E. Mogensen, Performance of VoIP on HSDPA, IEEE Vehicular Technology Conference, vol. 4, 30 May-1 June 2005, pp. 2335-2339.

[6] C. Boutremans and G. Iannaccone and C. Diot, Impact of link failures on VoIP performance, NOSSDAV 2002, pp. 63-71.

[7] D.T. Chen, N. Natarajan, and Y. Sun, “On the Simulation, Modeling, and Performance Analysis of an 802.16E Mobile Broadband Wireless Access System”, Communications and Computer Networks, 2005.

[8] R. G. Cole and J. H. Rosenbluth, “Voice over IP performance monitoring, Computer Communication Review, vol. 31, no. 2, 2001, pp. 9-24.

[9] L. Ding and R.A. Goubran, “Speech quality prediction in VoIP using the extended E-model,” IEEE GLOBECOM 2003, pp. 3974-3978.

[10] L. Hang, M. El Zarki, “Performance of H.263 video transmission over wireless channels using hybrid ARQ”, IEEE Journal on Selected Areas in Communications, vol. 15, Issue 9, Dec. 1997, pp. 1775-1786.

[11] M.C. Hui, H.S. Matthews, Comparative analysis of traditional telephone and voice-over-Internet protocol (VoIP) systems, IEEE International Symposium on Electronics and the Environment, May 2004, pp. 106-111.

[12] W. Jiang, H. Schulzrinne, “Comparison and optimization of packet loss repair methods on VoIP perceived quality under bursty loss”, International workshop on Network and operating systems support for digital audio and video, 2002, pp. 73-81.

[13] I. Joe, “An adaptive hybrid ARQ scheme with concatenated FEC codes for wireless ATM,” In Proc. of ACM/IEEE MobiCom 1997, pp. 131-138.

[14] Y. Li, A. Markopoulou, N. Bambos, and J. Apostolopoulos, “Joint Power/Playout Control Schemes for Media Streaming over Wireless Links”, In Proc. of IEEE Packet Video Workshop 2004, Dec. 2004.

[15] A. Majumdar, D.G. Sachs, I. Kozintsev, K. Ramchandran, M.M. Yeung, “Multicast and unicast real-time video streaming over wireless LANs”, IEEE Transactions on Circuits and Systems for Video Technology, vol. 12, Issue 6, June 2002, pp. 524-534.

[16] A.P. Markopoulou, F.A. Tobagi, M.J. Karam, “Assessment of VoIP quality over Internet backbones”, IEEE INFOCOM 2002. vol. 1, pp. 150-159.

[17] S. Perera, H. Sirisena, “Contention Based Negative Feedback ARQ for VoIP Services in IEEE 802.16 Networks”’, IEEE International Conference on Networks, 2006, vol. 2, Sept. 2006, pp. 1-6.

[18] R. Rajavelsamy, V. Jeedigunta, B. Holur, M. Choudhary, O. Song, Performance evaluation of VoIP over 3G-WLAN interworking system, IEEE Wireless Communications and Networking Conference (WCNC), vol. 4, pp. 2312-2317.

[19] D.G. Sachs, I. Kozintsev, M. Yeung, D.L Jones, “Hybrid ARQ for robust video streaming over wireless LANs”, International Conference on Information Technology: Coding and Computing, 2001, pp. 317-321.

[20] M. van der Schaar, S, Krishnamachari, S. Choi, X. Xu, “Adaptive cross-layer protection strategies for robust scalable video transmission over 802.11 WLANs”, IEEE Journal on Selected Areas in Communications, vol. 21, Issue 10, Dec. 2003, pp. 1752-1763.

[21] S. Sengupta, M. Chatterjee, S. Ganguly, and R. Izmailov, “Improving R-Score of VoIP Streams over WiMax”, IEEE International Conference On Communications (ICC), vol. 2, June, 2006, pp. 866-871.

[22] D. Triantafyllopoulou, N. Passas and A. Kaloxylos, “A Cross-Layer Optimization Mechanism for Multimedia Traffic over IEEE 802.16 Networks”, European Wireless, 2007.

[23] S.J. Vaughan-Nichols, “Achieving wireless broadband with WiMax”, IEEE Computer, vol. 37, Issue 6, June 2004, pp. 10-13.

[24] B. Vucetic, “An adaptive coding scheme for time-varying channels,” IEEE Transactions on Communications, vol. 39, Issue 5, May 1991, pp. 653-663.

[25] C. Young-June, B. Saewoong, Scheduling for VoIP service in cdma2000 1x EV-DO, IEEE International Conference on Communications (ICC), vol. 3, 2004, pp. 1495-1499.

[26] B. Sklar, Digital Comn., 2nd ed. Prentice Hall.

[27] http://www.wimaxforum.org, 2006.

[28] http://en.wikipedia.org/wiki/WiMAX