# Improving Quality of VoIP Streams over WiMax

Shamik Sengupta, *Student Member*, *IEEE*, 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. Although 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. The IEEE 802.16-based WiMax promises to be one of the 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 the construction and transmission of *MAC protocol data units* (MPDUs) for supporting multiple VoIP streams. We study the quality of VoIP calls, usually given by R-score, with respect to the 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 increase the R-score and mean opinion score by about 40 percent.

**Index Terms**—VoIP, R-score, WiMax, FEC, ARQ, aggregation, fragmentation.

## 1 Introduction

Despite the growing popularity of data services, voice services still remain the major revenue earner for network service providers. The two most popular ways of providing voice services are packet switched telephone networks (PSTNs) and wireless cellular networks. The deployment of both of these forms of networks requires infrastructures that are usually very expensive. Alternative solutions are being sought which 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 that are used to carry voice signals over the IP network are commonly referred to as 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. Although some codecs are capable of some levels of adaptation and error concealment, the 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 is becoming more essential as there is already a huge demand for real-time services over wireless networks. Although 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 far exceed the capacity of current 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 user, either a residential home or a business facility. Among the new wireless broadband access technologies that are being considered, worldwide interoperability of microwave access (WiMax) is perhaps the strongest contender that is being supported and developed by a consortium of companies [27].

### 1.1 Worldwide Interoperability of Microwave Access

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 a simplifying but also a 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 the 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 of delivering high-speed services up to a range of 30 miles, thus posing strong competition to the existing last mile broadband access technologies, such as cable and DSL. WiMax uses multiple channels for a single transmission and

- *S. Sengupta and M. Chatterjee are with the School of Electrical Engineering and Computer Science, University of Central Florida, 4000 Central Florida Blvd., Orlando, FL 32816. E-mail: {Shamik, mainak}@cpe.ucf.edu.*
- *S. Ganguly is with NEC Laboratories America, 4 Independence Way, Princeton, NJ 08540. E-mail: samrat@nec-labs.com.*

*Manuscript received 1 May 2006; revised 9 July 2007; accepted 16 July 2007; published online 22 Aug. 2007.*
*Recommended for acceptance by A. Zomaya.*
*For information on obtaining reprints of this article, please send e-mail to: tc@computer.org, and reference IEEECS Log Number TC-0167-0506.*
*Digital Object Identifier no. 10.1109/TC.2007.70804.*

provides bandwidth of up to 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 50 businesses with T1-level (1.544 Mbps) services and hundreds of homes with high-speed Internet access. WiMax's low cost of deployment coupled with existing demands from underserved areas creates major business opportunities.

### 1.2 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 listed as follows:

- We show how the quality of VoIP calls is represented by R-score, which primarily depends on the 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* (MPDUs) 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 (FEC) 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 the 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 increase the R-score by about 40 percent.

### 1.3 Organization

The rest of this paper is organized as follows: In Section 2, we provide a brief overview on the adaptive techniques that have been proposed to support data/streaming services over wireless channels. In Section 3, we discuss the rich set of MAC-layer features of WiMax, with particular emphasis on aggregation and fragmentation. In Section 4, we show the effect of delay and loss on R-score, which is 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 5. In Section 6, we present the simulation model and results. Conclusions are drawn in Section 7.

## 2 Related Work

Supporting real-time applications over any wireless network (for example, 3G cellular networks, IEEE 802.11-based wireless LANs, 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 which combine audio and video processing techniques with mechanisms that are usually dealt with in the data link and physical layers. These approaches can broadly be classified into two categories: automatic repeat request (ARQ) and forward error correction (FEC). ARQ schemes provide high reliability when the channel is good or moderate. However, for error-prone channels, the throughput drops due to the increased frequency of retransmissions. In order to counter this effect, hybrid ARQ schemes are used which combine the FEC with the 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 Multirate (AMR) voice codec was properly tuned for IEEE 802.16 networks that allowed switching to the maximum encoding rate [22]. Note that such codecs can also be tuned for other networks as well. The study in [7] presents a simulation model and analyzes the performance of an 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 QoS differentiation is not needed in 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. In [12], the effect of bursty packet losses in the Internet was taken care of by changing the packet interval. 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

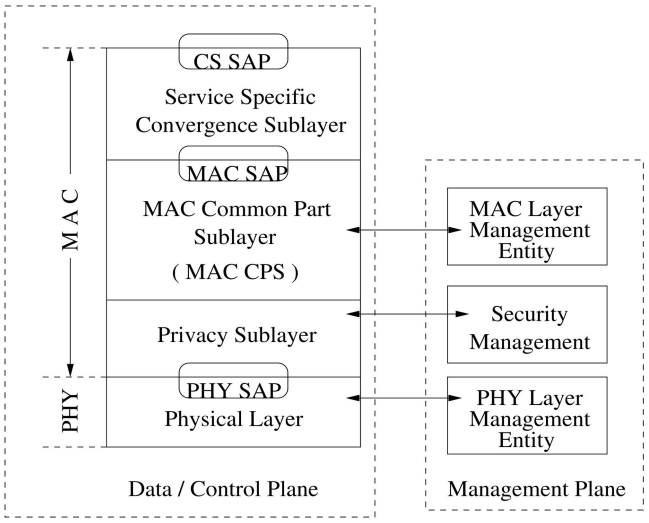



Fig. 1. WiMax MAC layer with SAPs.

rapid rerouting were 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 did not require changes in the underlying infrastructure. In [17], retransmission strategies for VoIP packets were deployed using unsolicited grant service (UGS) scheduling.

In this paper, we do not propose a new link-layer technique. Instead, 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.

## 3 The MAC Layer of WiMax

WiMax offers some flexible features that can potentially be exploited for delivering real-time services. In particular, although 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 is comprised of three sublayers which interact with each other through the service access points (SAPs), as shown in Fig. 1. The service-specific convergence sublayer provides the transformation or mapping of external network data with the help of the SAP. 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 MPDUs. The 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

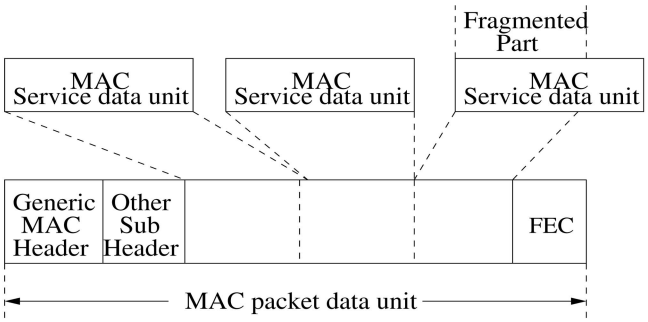



Fig. 2. MPDU accommodating multiple MSDUs.

connection identifier (CID) to identify which connection the MPDU is servicing.

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 this paper, we exploit this feature of the common part sublayer that modifies the size of the MPDUs to adapt to the varying channel conditions.

### 3.1 Aggregation

The common part sublayer is capable of packing more than one complete or partial MSDUs into one MPDU. In Fig. 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 the subscriber station.

### 3.2 Fragmentation

The common part sublayer can also fragment an MSDU into multiple MPDUs. In Fig. 3, we show how a portion of a single MSDU occupies the entire payload field of an 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.

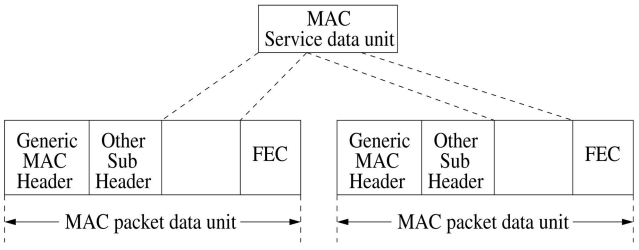



Fig. 3. Single MSDU fragmented to multiple MPDUs.

## 4 Delay and Loss Sensitivity of VoIP

As VoIP packets travel through a network, there evidently are some congestion and channel-related losses. In addition, the packets suffer delay, depending on the congestion at the intermediate routers. Both the loss and delay of packets adversely affect the quality of VoIP calls, which is generally expressed in terms of R-score.

### 4.1 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 depacketized and forwarded to a playout buffer, which smooths 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 is therefore measured by the MOS. The MOS is a subjective quality score that ranges from 1 (worst) to 5 (best) and is obtained by conducting subjective surveys. Although these methods provide a good assessment technique, they fail to provide an online assessment, which might be used for adaptation purposes. 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 their influence can only be measured in an isolated manner. The 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 yet been made and, for this reason, the 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 the MOS through the following nonlinear mapping [3]:

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

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

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

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

### 4.2 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, that is, the R-score.

#### 4.2.1 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 the 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 $x < 0$; otherwise, this is 1.

#### 4.2.2 Effect of Loss

The VoIP call quality is also dependent on the loss impairment. Recall that $I_e$ represents the effect of packet loss rate. $I_e$ accounts for impairments caused by both the network's and the 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], which relates $I_e$ to the overall packet loss rate as

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

where $\gamma_1$ is a constant that determines the voice quality impairment caused by encoding and $\gamma_2$ and $\gamma_3$ describe the impact of loss on the 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 the 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 the G.729a and G.711 codecs are shown in Table 1 [3].

#### 4.2.3 Sensitivity of R-Score toward Delay and Loss

We rewrite (3), using (5) and (6), as

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

TABLE 1
Loss Impairment Parameters [3]

| Codec | $\gamma_1$ | $\gamma_2$ | $\gamma_3$ |
|---|---|---|---|
| G.729a | 11 | 40 | 10 |
| G.711 | 0 | 30 | 15 |

To find the sensitivity of loss and delay toward the R-score of VoIP calls, we use different delay values, keeping the network loss fixed, in (8). In Fig. 4a, we observe how the R-score changes with increasing delay for fixed loss. Similarly, in Fig. 4b, we show how the R-score changes with loss for fixed delay. In Fig. 4a, we observe that there is no significant drop in the R-score for a given network loss rate. The R-score drops relatively rapidly when the delay exceeds 177.3 ms. Note that this is the delay threshold defined in [3]. However, when the loss is increased from 0 percent to 5 percent, the drop in the R-score is about 14. Similarly, in Fig. 4b, we see that the drop in the R-score is mainly due to the loss ($x$-axis). When the delay is significantly increased from 50 to 200 ms, the drop in the 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 limits.

## 5 Adaptive MPDU Construction

With the sensitivity of VoIP with respect to the 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.

### 5.1 Connection Setup 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 that provides 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 simultaneously handles multiple VoIP calls. 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 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, that is, the number, of VoIP streams.

#### 5.1.1 Phase 1: Subscriber Station Requests Connection Request

A 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-parameter requests (bandwidth, frequency, and peak or average rate) are sent next and variable length MSDU indicators are turned on.

#### 5.1.2 Phase 2: Base Station Confirms Connection

After receiving a 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.

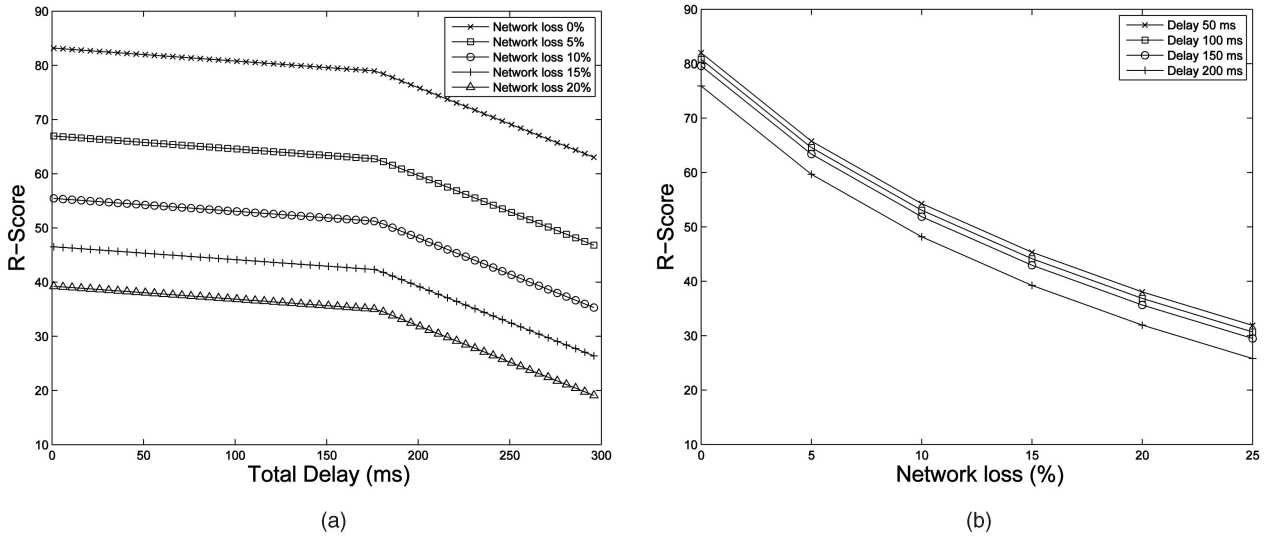



Fig. 4. (a) R-score versus delay. (b) R-score versus network loss.

#### 5.1.3 Phase 3: Base Station Starts Transmission of MPDUs

The MSDUs obtained from the MAC-convergence sublayer are converted to MPDUs. As needed, MSDUs can be either packed or fragmented to form the desired sized MPDUs. Since no feedback is received at the start of transmission, the payload and code sizes agreed upon at the time of connection establishment are maintained. When feedback is received, the next awaiting MPDU is formed, depending on the type of feedback received. On the reception of feedback, 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.

### 5.2 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 FEC 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, that is, *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 an MPDU, the resulting bit loss probability is given by [26]

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

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

#### 5.2.1 Decrease Payload, with Code Size Fixed

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

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

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

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

The ratio of $p'$ to $p$ is given by

$$\frac{p'}{p} = \frac{(1-b')^{(M+N')}}{(1-b)^{(M+N)}} = \left(\frac{1-b'}{1-b}\right)^{(M+N)} \times \frac{1}{(1-b')^{(N-N')}}. \qquad (12)$$

As $b' < b$, $(1-b') > (1-b)$. Thus, the first term in (12) $(\frac{1-b'}{1-b})^{(M+N)}$, , is greater than 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 $p' > p$.

#### 5.2.2 Increase Code Size with Payload Fixed

Similarly, it can be argued that if the code is increased while keeping the payload fixed, then the resulting bit loss probability decreases and the packet restore probability of MPDUs increases.

#### 5.2.3 Increase Both Payload and Code Size

The third scheme would be to increase both the payload and the code sizes. As we know, increasing the payload only will increase the resulting BER, so we must also increase the code to compensate for the increased payload.

### 5.3 Enabling the ARQ Mechanism

Although the application of FEC enhances the packet restore probability, the performance can still be further improved if the optional 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 setup. 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 feedback in terms of the packet restore probability, from which the WiMax MAC common part sublayer gets the information if a packet has been successfully received or not. In addition, this feedback gives an estimate about the channel status.

We apply fast feedback at the MAC layer and use very small packets to reduce the overhead. The parameters used in the feedback packets are the CID, the ARQ status (enabled or disabled), the maximum retransmission limit, the 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 correctly received, that is, the packet restore probability is below a certain threshold, then a retransmission mechanism is applied. The main advantage of using the retransmission scheme is that this lowers 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 correctly restored 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 by (4), is modified as

$$d = d_{codec} + d_{playout} + d_{network} + d_{queue}. \qquad (13)$$

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

### 5.4 Optimal MPDU Size

Since packets often get lost or corrupted during transmission in error-prone wireless channels, the 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 when the last bit of an MPDU is transmitted and the time when the acknowledgment of that MPDU is received. Moreover, we assume a zero time interval between the transmissions of two consecutive MPDUs, that is, the last bit of an 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 an MPDU is large, then the transmission time is large but the overhead due to headers is less, which helps in maintaining a high goodput. If an MPDU is dropped or corrupted due to a 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, that is, 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 the 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 the low payload/overhead ratio. Thus, we observe that both large and small MPDUs have their advantages and disadvantages. We propose combining 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 pseudocode for the adaptive MPDU construction is presented in Algorithm 1 ($n_1$, $m_1$, $m_2$, $n_3$, and $m_3$ are implementation-dependent parameters (see Table 2)):

**Algorithm 1**: feedback-based adaptive MPDU construction.
Input: feedback classification (Table 2)
For each transmitted MPDU $i$ {
BS receives a feedback from the receiver

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

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

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

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

TABLE 2
Feedback Classification

| 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 |

### 5.5 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 Mbps symbol rate, there are 5,000 physical symbols within a 1 ms frame.

In addition to the already proposed mechanisms, we propose another key aspect of the dynamic minislot allocation for not only enhancing the performance of the VoIP calls but also accommodating more numbers of calls. For the G.729a codec, a typical VoIP packet is 60 bytes (40 byte RTP/UDP/IP header and 20 byte 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 the number of retransmissions of this frame and the codec efficiency) are appended for error recovery. Thus, transmission of an MPDU (consisting of a single MSDU) takes about 8-10 $\mu$s. On the other hand, the minimum and maximum minislot duration are 1 physical symbol (0.2 $\mu$s) and 128 physical symbols ($\approx 26$ $\mu$s), respectively, with a 20 Mbps symbol rate. Thus, the duration of minislot allocated plays a vital role for VoIP packets. If a single minislot with a duration of 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 wasting minislots. Now, the question that arises is 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, 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 is such that the duration of all the combined minislots is greater than or equal to the MPDU(s) being transmitted.

## 6 Simulation Model and Results

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

### 6.1 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 BER: The *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.

### 6.2 Simulation Parameters for VoIP

For our simulation, we assumed that the VoIP streams were generated by the G.729a codec. Note that we could have used the specifications of any other voice codec (for example, 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 Table 3.

TABLE 3
Simulation Parameters

| 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 |

### 6.3 Simulation Results for VoIP

In Figs. 5a and 5b, we present the packet restore probability for both the nonadaptive and adaptive schemes. For this simulation, we assumed that the channel remains in the good, medium, and bad states for 30 percent, 50 percent, and 20 percent of the time, respectively.

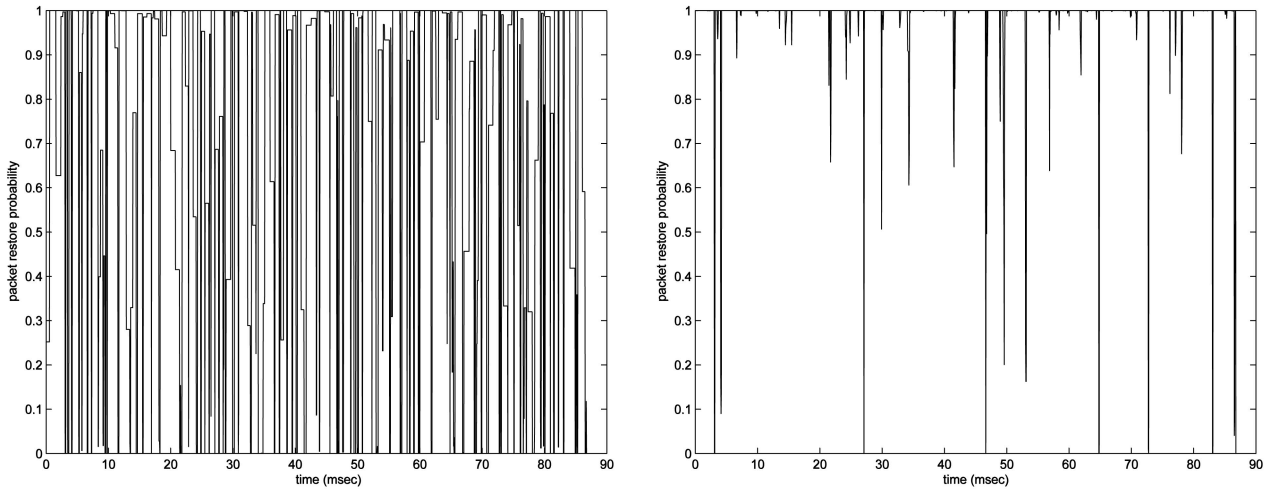
In the adaptive scheme, we used both the ARQ and aggregation schemes. In the nonadaptive scheme, we disabled the ARQ mechanism. With the adaptive scheme, the packet restore probability is significantly improved.

In Fig. 6a, we present the R-score for both the adaptive and nonadaptive schemes without competing traffic. It is seen that, with the adaptive scheme, there is an improvement of about 40 percent in the R-score, which indicates that the call quality can be increased in WiMax by using aggregation and ARQ on top of the MAC common part sublayer. It can also be noted that, with 2,000 streams, the R-score is still above 70, whereas, with 1,500 streams, it is above 73.

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

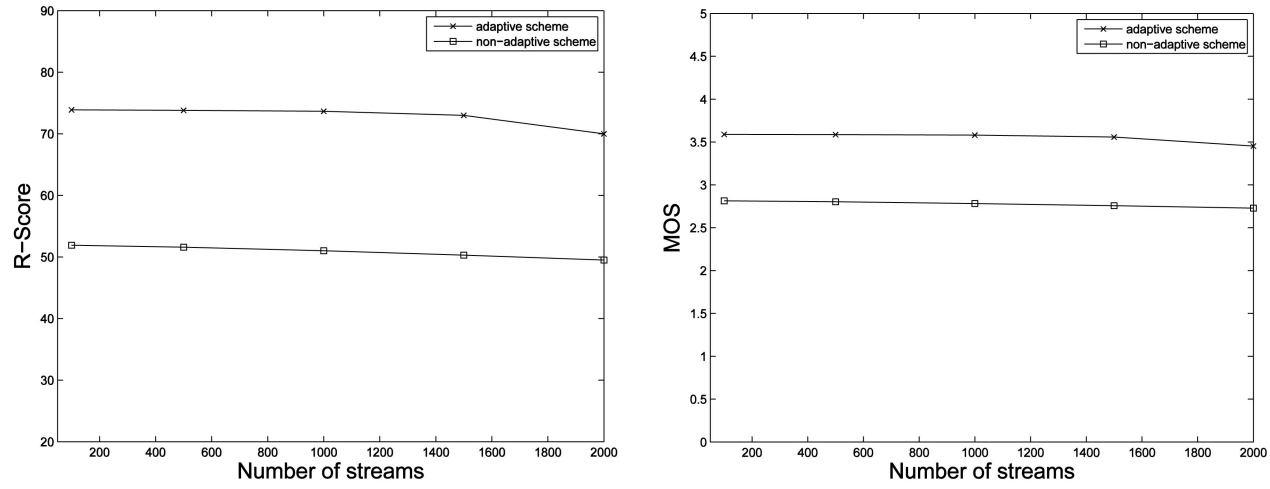
We have also modified our simulation model and introduced intermittent data sessions which generate competing traffic for the VoIP streams. The additional plots with competing traffic are shown in Figs. 7a and 7b. We observe that, even with intermittent data traffic, the R-score and MOS for the adaptive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

Next, we varied the maximum number of retransmissions from 1 to 3. It can be noted that, even with the allowed number of retransmissions, a packet might not be restored.



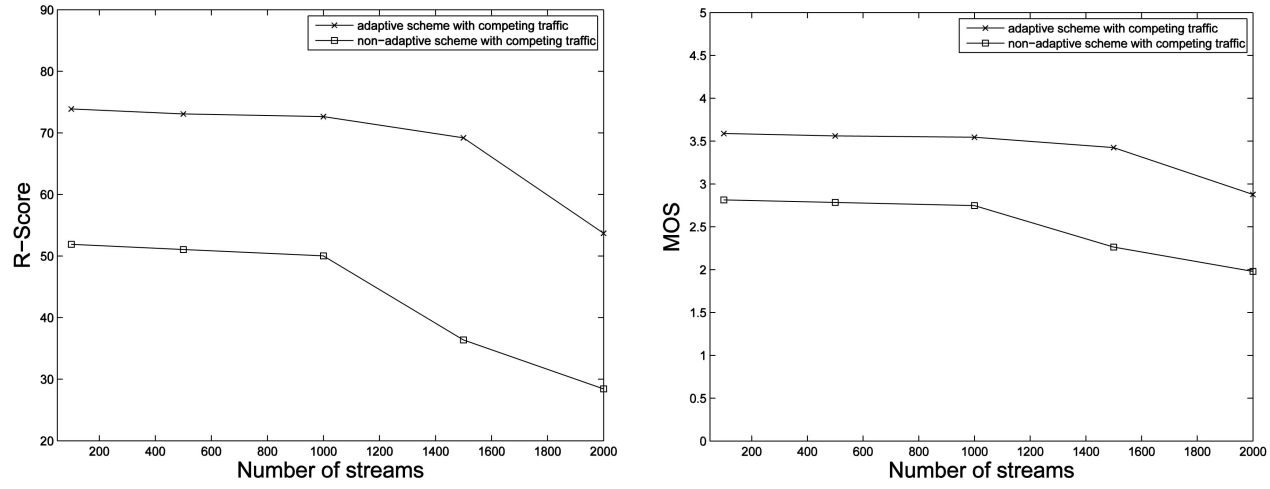

(a) (b)

Fig. 5. (a) Packet restore probability for the nonadaptive scheme. (b) Packet restore probability for the adaptive scheme.




(a) (b)

Fig. 6. (a) R-score versus the number of VoIP streams. (b) MOS versus the number of VoIP streams.




(a) (b)

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

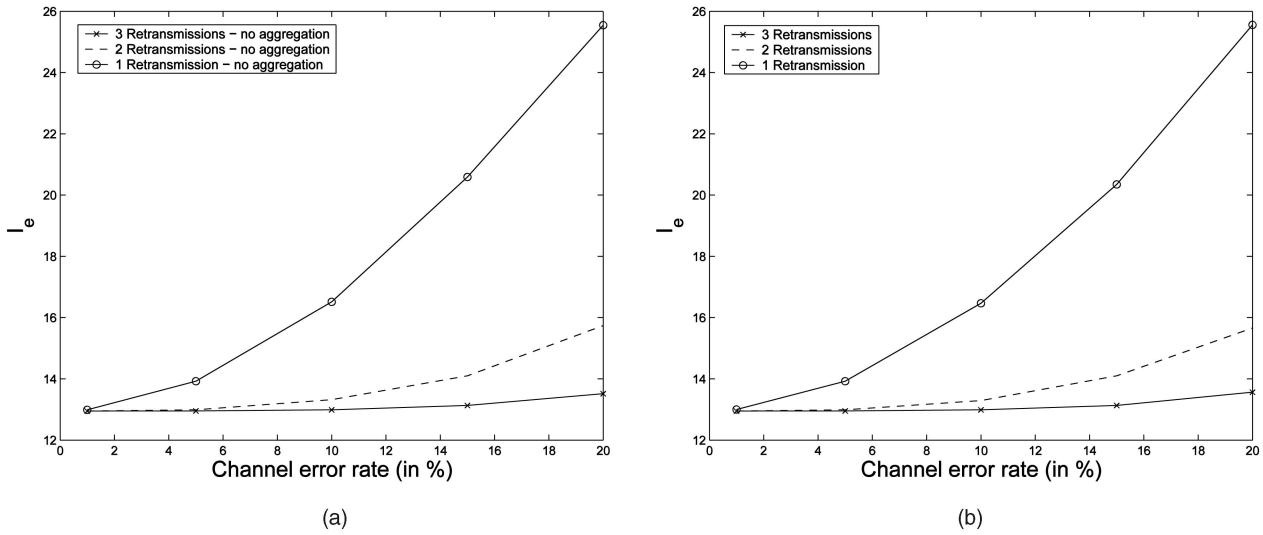



Fig. 8. (a) Loss impairment without aggregation. (b) Loss impairment with aggregation.

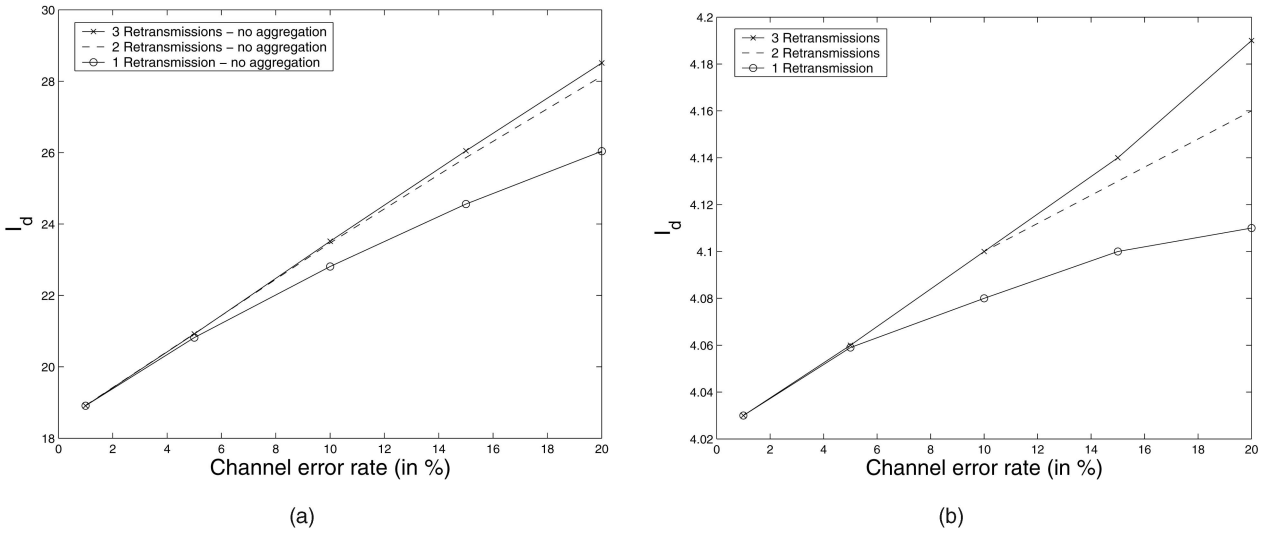



Fig. 9. (a) Delay impairment without aggregation. (b) Delay impairment with aggregation.

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 1,000 and gradually increased the channel error rate. Retransmission with aggregation and retransmission without aggregation are studied separately.

It is shown in Figs. 8a and 8b 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 shown in Figs. 9a and 9b (note the difference in the range of the $y$-axes in Figs. 9a and 9b). In Figs. 8b and 9b, we find that the loss impairment is greater with the increase in the channel error rate than the delay impairment for the retransmission with aggregation. On the other hand, for the retransmission without aggregation, the delay impairment is greater than the loss impairment. This is because the requested minislot(s) are not fully utilized when packets are retransmitted without aggregation.

In Figs. 10a and 10b, we present the variation of the R-score and MOS versus the channel error rate with and without aggregation. It is evident in Figs. 10a and 10b that there is an improvement in the R-score and MOS, particularly when the allowed numbers of retransmissions are 2 and 3. It is observed that, with the increase in channel error rates, the rate of decrease is much less for two or three retransmissions than for just one retransmission. It is also evident that the retransmission with the aggregation scheme gives better R-score and MOS values than the retransmission without aggregation. Thus, it is desirable to bundle both features (retransmission and aggregation) in WiMax to improve the call quality in VoIP.

In Figs. 11a and 11b, we show how the R-score is affected when the number of streams is increased. The channel error rate is assumed to be 20 percent. As expected, the retransmission coupled with aggregation yields a better R-score. In Fig. 11a, the R-score is presented without aggregation, whereas, in Fig. 11b, the R-score is presented with aggregation. Aggregation helps in putting multiple

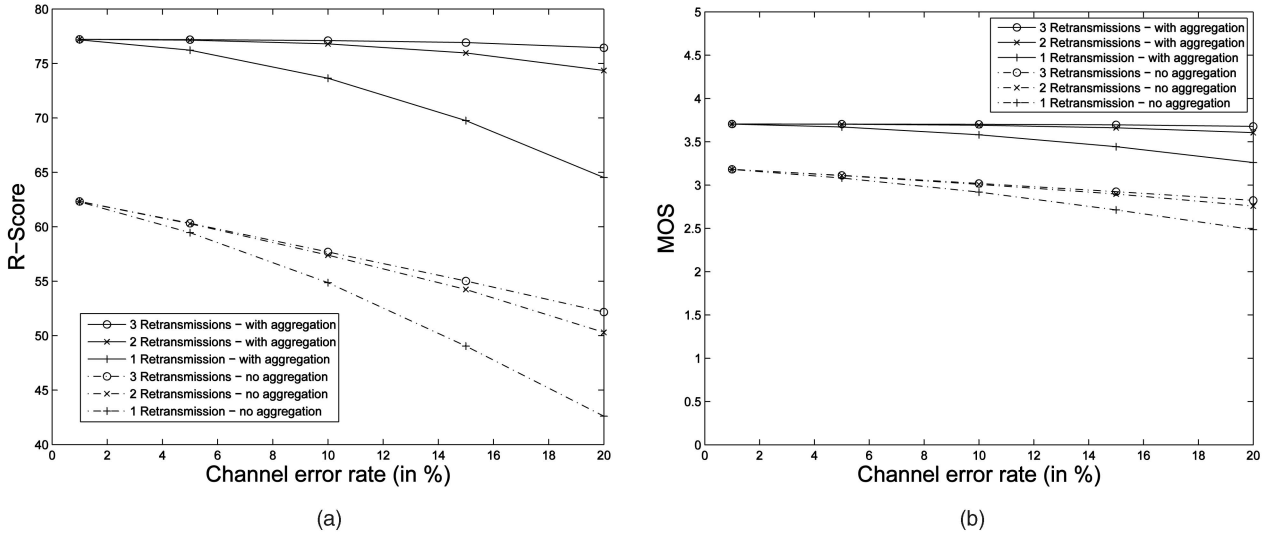



Fig. 10. (a) R-score versus error rate with and without aggregation. (b) MOS versus error rate with and without aggregation.

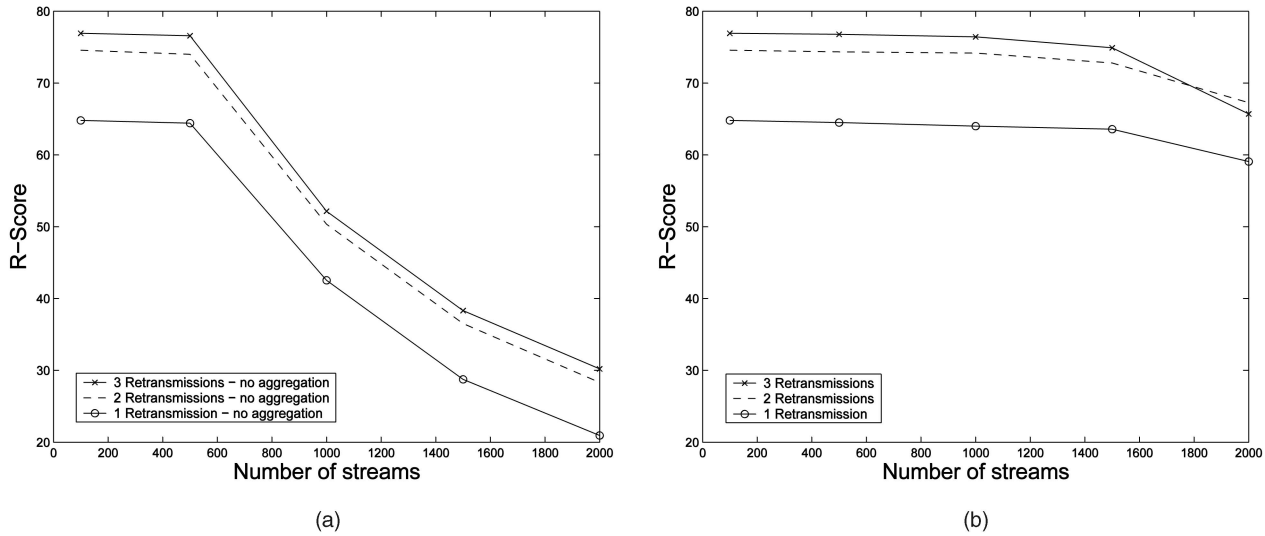



Fig. 11. (a) R-score versus the number of streams without aggregation. (b) R-score versus the number of streams with aggregation.

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 greater numbers of VoIP streams without affecting their call quality.

Moreover, it is seen that, with the retransmission with the aggregation scheme, the three-retransmission scheme gives a better performance for low and medium load than the two and one-retransmission schemes, but the performance degrades with an increase in the number of streams. The reason behind this is that, in a typical cell, the VoIP streams share the common backhaul bandwidth. With greater numbers of streams, the bandwidth allocation for each stream decreases. On the other hand, with an increase in the number of retransmissions and the 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 the delay impairment. As the R-score decreases with the increasing delay impairment, we find that the three-retransmission scheme produces worse call quality with increasing load.

## 7 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 the extension of VoIP calls over wireless networks is 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 greater numbers of VoIP calls. The proposed schemes make use of the flexible MAC features, in particular 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 is studied with respect to the R-score. We exploit the difference in sensitivity of the R-score toward loss and delay for recovering as many packets as possible, at the cost of increased delay. Exhaustive simulation experiments reveal that the feedback-based techniques coupled with retransmissions, aggregation, and variable-sized MPDUs increase not only the R-score (and, consequently, the MOS) but also the number of VoIP streams.

## ACKNOWLEDGMENTS

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

## REFERENCES

[1] *IEEE Standard for Local and Metropolitan Area Networks Part 16: Air Interface for Fixed Broadband Wireless Access Systems,* IEEE Standard 802.16-2001 Working Group Std., 2002.
[2] *IEEE Standard for Local and Metropolitan Area Networks Part 16: Air Interface for Fixed Broadband Wireless Access Systems—Amendment 2: MAC Modifications and Additional Physical Layer Specifications for 2-11 GHz Standard 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,* Dec. 1998.
[4] Y. Amir, C. Danilov, S. Goose, D. Hedqvist, and A. Terzis, "An Overlay Architecture for High-Quality VoIP Streams," *IEEE Trans. Multimedia,* vol. 8, no. 6, pp. 1250-1262, Dec. 2006.
[5] W. Bang, K.I. Pedersen, T.E. Kolding, and P.E. Mogensen, "Performance of VoIP on HSDPA," *Proc. 61st IEEE Vehicular Technology Conf.,* vol. 4, pp. 2335-2339, 2005.
[6] C. Boutremans, G. Iannaccone, and C. Diot, "Impact of Link Failures on VoIP Performance," *Proc. 12th Int'l Workshop Network and Operating Systems Support for Digital Audio and Video,* pp. 63-71, 2002.
[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," *Comm. and Computer Networks,* 2005.
[8] R.G. Cole and J.H. Rosenbluth, "Voice over IP Performance Monitoring," *Computer Comm. Rev.,* vol. 31, no. 2, pp. 9-24, 2001.
[9] L. Ding and R.A. Goubran, "Speech Quality Prediction in VoIP Using the Extended E-Model," *Proc. IEEE Global Telecomm. Conf.,* pp. 3974-3978, 2003.
[10] L. Hang and M. El Zarki, "Performance of H.263 Video Transmission over Wireless Channels Using Hybrid ARQ," *IEEE J. Selected Areas in Comm.,* vol. 15, no. 9, pp. 1775-1786, Dec. 1997.
[11] M.C. Hui and H.S. Matthews, "Comparative Analysis of Traditional Telephone and Voice-over-Internet Protocol (VoIP) Systems," *Proc. IEEE Int'l Symp. Electronics and the Environment,* pp. 106-111, May 2004.
[12] W. Jiang and H. Schulzrinne, "Comparison and Optimization of Packet Loss Repair Methods on VoIP Perceived Quality under Bursty Loss," *Proc. 12th Int'l Workshop Network and Operating Systems Support for Digital Audio and Video,* pp. 73-81, 2002.
[13] I. Joe, "An Adaptive Hybrid ARQ Scheme with Concatenated FEC Codes for Wireless ATM," *Proc. ACM/IEEE MobiCom '97,* pp. 131-138, 1997.
[14] Y. Li, A. Markopoulou, N. Bambos, and J. Apostolopoulos, "Joint Power/Playout Control Schemes for Media Streaming over Wireless Links," *Proc. 14th IEEE Int'l Packet Video Workshop,* Dec. 2004.
[15] A. Majumdar, D.G. Sachs, I. Kozintsev, K. Ramchandran, and M.M. Yeung, "Multicast and Unicast Real-Time Video Streaming over Wireless LANs," *IEEE Trans. Circuits and Systems for Video Technology,* vol. 12, no. 6, pp. 524-534, June 2002.
[16] A.P. Markopoulou, F.A. Tobagi, and M.J. Karam, "Assessment of VoIP Quality over Internet Backbones," *Proc. IEEE INFOCOM '02,* vol. 1, pp. 150-159, 2002.
[17] S. Perera and H. Sirisena, "Contention-Based Negative Feedback ARQ for VoIP Services in IEEE 802.16 Networks," *Proc. 14th IEEE Int'l Conf. Networks,* vol. 2, pp. 1-6, Sept. 2006.
[18] R. Rajavelsamy, V. Jeedigunta, B. Holur, M. Choudhary, and O. Song, "Performance Evaluation of VoIP over 3G-WLAN Interworking System," *Proc. IEEE Wireless Comm. and Networking Conf.,* vol. 4, pp. 2312-2317, 2005.
[19] D.G. Sachs, I. Kozintsev, M. Yeung, and D.L. Jones, "Hybrid ARQ for Robust Video Streaming over Wireless LANs," *Proc. Int'l Conf. Information Technology: Coding and Computing '01,* pp. 317-321, 2001.
[20] M. van der Schaar, S. Krishnamachari, S. Choi, and X. Xu, "Adaptive Cross-Layer Protection Strategies for Robust Scalable Video Transmission over 802.11 WLANs," *IEEE J. Selected Areas in Comm.,* vol. 21, no. 10, pp. 1752-1763, Dec. 2003.
[21] S. Sengupta, M. Chatterjee, S. Ganguly, and R. Izmailov, "Improving R-Score of VoIP Streams over WiMax," *Proc. IEEE Int'l Conf. Comm.,* vol. 2, pp. 866-871, June 2006.
[22] D. Triantafyllopoulou, N. Passas, and A. Kaloxylos, "A Cross-Layer Optimization Mechanism for Multimedia Traffic over IEEE 802.16 Networks," *Proc. 13th European Wireless Conf.,* 2007.
[23] S.J. Vaughan-Nichols, "Achieving Wireless Broadband with WiMax," *Computer,* vol. 37, no. 6, pp. 10-13, June 2004.
[24] B. Vucetic, "An Adaptive Coding Scheme for Time-Varying Channels," *IEEE Trans. Comm.,* vol. 39, no. 5, pp. 653-663, May 1991.
[25] C. Young-June and B. Saewoong, "Scheduling for VoIP Service in CDMA2000 1x EV-DO," *Proc. IEEE Int'l Conf. Comm.,* vol. 3, pp. 1495-1499, 2004.
[26] B. Sklar, *Digital Communications,* second ed. Prentice Hall, 2001.
[27] http://www.wimaxforum.org, 2006.
[28] http://en.wikipedia.org/wiki/WiMAX, 2007.

**Shamik Sengupta** received the BE degree (with first class honors) in computer science and engineering from Jadavpur University, Calcutta, in 2002. He is currently working toward the PhD degree in the School of Electrical Engineering and Computer Science at the University of Central Florida. He was with Tata Consultancy Services, India, as an assistant systems engineer. His research interests include resource management in wireless networks, auction and game theories, pricing, and WMAN technologies. He is a student member of the IEEE.

**Mainak Chatterjee** received the BSc (Hons) degree in physics from the University of Calcutta in 1994, the ME degree in electrical communications and engineering from the Indian Institute of Science, Bangalore, in 1998, and the PhD degree from the University of Texas at Arlington in 2002. He is currently an assistant professor in the School of Electrical Engineering and Computer Science at the University of Central Florida. He is on the executive and technical program committees of several international conferences. His research interests include economic issues in wireless networks, applied game theory, resource management and quality-of-service provisioning, ad hoc and sensor networks, CDMA data networking, and link-layer protocols.

**Samrat Ganguly** received the BSc degree in physics from the Indian Institute of Technology, Kharagpur, India, in 1994, the ME degree in computer science from the Indian Institute Science, Bangalore, India, in 1998, and the PhD degree in computer science from Rutgers University, Piscataway, New Jersey, in 2003. Since 2001, he has been a research staff member at NEC Laboratories America, Princeton, New Jersey. His research interests include distributed algorithm designs and performance optimization in wireless, overlay, and content delivery networks.