Scalable Video Transport over Wireless IP Networks

Dapeng Wu* Yiwei Thomas Hou† Ya-Qin Zhang§

* Carnegie Mellon University, Pittsburgh, PA, USA
† Fujitsu Laboratories of America, Sunnyvale, CA, USA
§ Microsoft Research, Beijing, China

(Invited Paper)

Abstract

Recently, there has been great interest in transporting real-time video over wireless IP networks from both industry and academia. Real-time video applications have quality-of-service (QoS) requirements. However, fluctuations of wireless channel conditions pose many challenges to provide QoS for video transmission over wireless IP networks. It has been shown that scalable video coding and adaptive services are viable solutions under time-varying wireless environment. In this paper, we propose an adaptive framework to support quality video communication over wireless IP networks. The adaptive framework includes: (1) scalable video representations, (2) network-aware video applications, and (3) adaptive services. Under this framework, as wireless channel conditions change, the mobile terminal and network elements can scale the video streams and transport the scaled video streams to receivers with acceptable perceptual quality. The key advantages of the adaptive framework are: (1) perceptual quality is degraded gracefully under severe channel conditions; (2) network resources are efficiently utilized; and (3) the resources are shared in a fair manner.

1 Introduction

The proliferation of multimedia on the World Wide Web and the emergence of broadband wireless networks have brought great interest in real-time video communications over wireless IP networks. However, delivering quality video over wireless networks in real-time is a challenging task. This is primarily because of the following problems.

Bandwidth fluctuations: First, the throughput of a wireless channel may be reduced due to multipath fading, co-channel interference, and noise disturbances. Second, the capacity of a wireless channel may fluctuate with the changing distance between the base station and the mobile host. Third, when a mobile terminal moves between different networks (e.g., from wireless local area network to wireless wide area network), the available bandwidth may vary drastically (e.g., from a few megabits per second to a few kilobits per second). Finally, when a handoff takes place, a base station may not have enough unused radio resource to meet the demand of a newly joined mobile host. As a result, bandwidth fluctuations is a serious problem for real-time video transmission over wireless networks.

High bit error rate: Compared with the wired links, wireless channels are typically much more noisy and have both small-scale (multipath) and large-scale (shadowing) fades, making the bit error rate (BER) very high. The resulting bit errors can have devastating effect on video presentation quality. Therefore, there is a critical need for robust transmission of video over wireless channels.

Heterogeneity: In multicast scenario, receivers may have different requirements and properties in terms of latency, visual quality, processing capabilities, power limitations (wireless vs. wired) and bandwidth limitations. The heterogeneous nature of receivers' requirements and properties make it difficult to design an efficient multicast mechanism.

It has been shown that scalable video is capable of coping with the variability of bandwidth gracefully [2, 12]. A scalable video coding scheme is to produce a compressed bit-stream, parts of which are decodable. Compared with decoding the complete bit-stream, decoding part of the compressed

0-7803-6465-5/00 $10.00

bit-stream produces pictures with degraded quality, or smaller image size, or smaller frame rate [7]. In contrast, non-scalable video is more susceptible to bandwidth fluctuations since it cannot adapt its video representation to bandwidth variations [12]. Thus, scalable video is more suitable for use in a wireless environment to cope with the fluctuation of wireless channels. Furthermore, scalable video representation is a good solution to heterogeneity problem in multicast case [12].

Recently, application-aware adaptive services have been demonstrated to be able to effectively mitigate fluctuations of resource availability in wireless networks [2]. Scalable video representation naturally fit unequal error protection, which can effectively combat bit errors induced by the wireless medium. This motivates us to present an adaptive framework to support quality video communication over wireless IP networks.

Our proposed adaptive framework consists of (1) scalable video representations, each of which has its own specified QoS requirement, (2) network-aware applications, which are aware of network status, and (3) adaptive services, which make network elements support the QoS requirements of scalable video representations. Under this framework, as wireless channel conditions change, the mobile terminal and network elements can scale the video streams and transport the scaled video streams to receivers with acceptable perceptual quality. Our adaptive framework has the following key features.

1. *Graceful quality degradation:* Different from non-scalable video, scalable video can adapt its video representation to bandwidth variations and the network can drop packets with awareness of the video representations. As a result, perceptual quality is gracefully degraded under severe channel conditions.

2. *Efficiency:* When there is excess bandwidth (excluding reserved bandwidth), the excess bandwidth will be efficiently used in a way that maximizes the perceptual quality or revenue.

3. *Fairness:* The resources can be shared in either a utility-fair manner [5] or a max-min fair manner.

Previous works include Naghshineh's adaptive framework [13], Lu's adaptive service [11], and Bianchi's adaptive services [5]. Our adaptive framework is different from these previous works in two aspects: (1) network-aware applications and (2) a scheduling architecture for scalable video streams, which achieves both QoS (bounded delay and throughput guarantee) and fairness.

The remainder of this paper is organized as follows. Section 2 describes network-aware applications. In Section 3, we present the adaptive services for transporting scalable video over wireless IP networks. Section 4 summarizes this paper and points out future research directions.

2 Network-aware Applications

The use of network-aware applications is motivated by the following facts: (1) the bit error rate is very high when the channel status is poor, and (2) packet loss is unavoidable if the available bandwidth is less than required. If a sender attempts to transmit each layer without any awareness of the channel status, all layers may get corrupted with equal probability, resulting in very poor picture quality. To address this problem, we propose to use network-aware applications, which preemptively discard enhancement layers at the sender in an intelligent manner by considering network status.

For the purpose of illustration, we present an architecture including a network-aware mobile sender, an application-aware base station, and a receiver in Fig. 1. The architecture in Fig. 1 is applicable to both live and stored video. In Fig. 1, at the sender side, the compressed video bit-stream is first filtered by the scaler, the operation of which is to select certain video layers to transmit. Then the selected video representation is passed through transport protocols. Before being transmitted to the base station, the bit-stream has to be modulated by a modem (i.e., modulator/demodulator). Upon receipt of the video packets, the base station transmits them to the destination through the Internet.

Note that a scaler can distinguish the video layers and drop layers according to their significance. The dropping order is from the highest enhancement layer down to the base layer. A scaler only performs two operations: (1) scale down the received video representation, that is, drop the enhancement layer(s); (2) transmit what is received, i.e., do not scale the received video representation.

Under our architecture, a bandwidth manager is maintained in the base station. One function of the bandwidth manager is to notify the sender about the available bandwidth of the wireless channel through signaling channel [14]. Upon receiv-

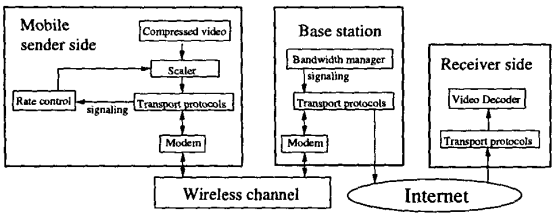

Figure 1: An architecture for transporting scalable video from a mobile terminal to a wired terminal.

ing this information, the rate control module at the sender conveys the bandwidth parameter to the scaler. Then, the scaler regulates the output rate of the video stream so that the transmission rate is less than or equal to the available bandwidth.

Another scenario is that the base station notifies the sender about the channel quality (i.e., BER) [3]. Upon receiving this information, the rate control module at the sender commands the scaler to perform the following (suppose that the video is compressed into two layers): (1) if the BER is above a threshold, discard the enhancement layer so that the bandwidth allocated for the enhancement layer can be utilized by forward error correction (FEC) to protect the base layer; (2) otherwise transmit both layers.

The network-aware application has two advantages. Firstly, by taking the available bandwidth into account, the sender can make the best use of network resources by selectively discarding enhancement layers in order to minimize the likelihood of more significant layers being corrupted, thereby increasing the perceptual quality of the video delivered. Secondly, by considering the channel error status, the sender can discard the enhancement layers and FEC can utilize the bandwidth allocated for the enhancement layer to protect the base layer, thereby maximizing the possibility of the base layer being correctly received.

Note that adaptive techniques at physical/link layer are required to support network-aware applications. Such adaptive techniques include a combination of variable spreading, coding, and code aggregation in Code Division Multiple Access (CDMA) systems, adaptive coding and modulation in Time Division Multiple Access (TDMA) systems, channel quality estimation, and measurement feedback channel [14]. In addition, the feedback interval is typically constrained on the order of tens to hundreds of milliseconds [14].

3 Adaptive Service

A scalable video encoder can generate multiple layers or substreams to the network. The adaptive service is to provide scaling of the substreams based on the resource availability conditions in the fixed and wireless network. Specifically, the proposed adaptive service includes the following functions.

- Reserve a minimum bandwidth to meet the demand of the base layer. As a result, the perceptual quality can always be achieved at an acceptable level.

- Adapt the enhance layers based on the available bandwidth and the fair policy. In other words, it scales the video streams based on resource availability.

Advantages of using scaling inside the network include:

(1) *Adaptiveness to network heterogeneity.* For example, when an upstream link with larger bandwidth feeds a downstream link with smaller bandwidth, use of a scaler at the connection point could help improve the video quality. This is because it selectively drops substreams instead of randomly dropping.

(2) *Low latency and low complexity.* Scalable video representations make the operation at a scaler very simple, i.e., only discarding enhancement layers. Thus, the processing is fast, compared with processing on non-scalable video.

(3) *Lower call blocking and handoff dropping probability.* The adaptability of scalable video at base stations can translate into lower call blocking prabability and handoff dropping probability.

The adaptive service can be deployed in the whole network (i.e., end-to-end provisioning) or only at base stations (i.e., local provisioning). Since local provisioning of the adaptive service is just a subset of end-to-end provisioning, we will focus on end-to-end provisioning in this paper.

The required components of the end-to-end adaptive service include: (1) service contract, (2) call admission control and resource reservation, (3) mobile multicast mechanism, (4) substream scaling, (5) substream scheduling, and (6) link-layer error control, which are described in Section 3.1 to 3.6, respectively.

3.1 Service contract

The service contract between the application and the network could consist of multiple subcontracts, each of which corresponds to one or more substreams with similar QoS guarantees. Each subcontract has to specify traffic characteristics and QoS requirements of the corresponding substream(s). A typical scenario is that a subcontract for the base layer specifies reserved bandwidth while a subcontract for the enhancement layers does not specify any QoS guarantee. As examples, we will use this typical scenario for two-layered video in the rest of the paper.

At a video source, substreams must be generated according to subcontracts used by the application and shaped at the network access point. In addition, a substream is assigned a priority according to its significance. For example, the base layer is assigned the highest priority. The priority can be used by routing, scheduling, scaling, and error control components of the adaptive network.

3.2 Call admission control and resource reservation

Call admission control (CAC) and resource reservation are two of the major components in end-to-end QoS provisioning. The function of CAC is to check whether admitting the incoming connection would reduce the service quality of existing connections, and whether the incoming connection's QoS requirements can be met. If a connection request is accepted, resources need to be reserved for this connection in two parts. First of all, in order to maintain the specified QoS in long time-scale, the network must reserve some resources along the current path of a mobile connection. Second, in order to seamlessly achieve the QoS at short time-scale, some duplication must be done in the transport of the connection to neighboring base stations of a connection so that in the event of a handoff, an outage in the link can be avoided.

The scalable video representation (i.e., substream) concept provides a very flexible and efficient solution to the problem of CAC and resource reservation. First, there is no need to reserve bandwidth for the complete stream since typically only base-layer substream needs QoS guarantee. As a result, CAC is only based on the requirement of the base layer and resource is reserved only for the base-layer substream. Second, the enhancement layer substream(s) of one connection could share the leftover bandwidth with the enhancement-layer substreams of other connections. The enhancement-layer substreams are subject to scaling under bandwidth shortage and/or severe error conditions, which will be discussed in Section 3.4.

3.3 Mobile multicast mechanism

CAC and resource reservation can provide connection-level QoS guarantee. To seamless guarantee QoS at packet level, mobile multicast mechanism has to be used. That is, while being transported along its current path, the base-layer stream is also multicast to its neighboring base stations so that QoS in small time-scale can be seamlessly achieved.

To support seamless QoS, the mobile routing protocol needs to be proactive and anticipatory in order to match the delay, loss, and jitter constraints of a substream. According to the requirements of a substream, multicast paths might need to be established. The multicast paths terminate at base stations that are potential access-point candidates of a mobile terminal. The coverage of such a multicast path depends on the QoS requirements and the mobility as well as handoff characteristics of a mobile receiver. As a mobile station hands off from a base station to another, new paths are added and old paths are deleted [13].

3.4 Substream scaling

Scaling is employed during bandwidth fluctuations and/or under poor channel conditions. As the available bandwidth on a path reduces due to mobility or fading, lower-priority substreams are dropped by the scaler(s) on the path and substreams with higher priority are transmitted. As more bandwidth becomes available, lower-priority substreams are passed through the scaler, and the perceptual quality at the receivers increases. Figure 1 showed an architecture for transporting scalable video from a mobile terminal to a wired terminal. Figure 2 depicts an architecture for transporting scalable video from a wired terminal to a mobile terminal. We do not show the case of transporting scalable video from a mobile terminal to a mobile terminal since it is a combination of Fig. 1 and Fig. 2.

The scaling decision is made by a bandwidth manager. When there is no excess bandwidth (excluding reserved bandwidth), the bandwidth manager instructs the scaler to drop the enhancement layers. If there is excess bandwidth, the excess bandwidth can be shared in either a utility-fair manner [5] or a max-min fair manner [11].

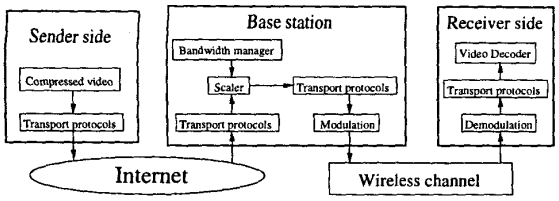

Figure 2: An architecture for transporting scalable video from a wired terminal to a mobile terminal.

3.5 Substream scheduling

The substream scheduler is used in mobile terminals as well as base stations. Its function is to schedule the transmission of packets on the wireless medium according to their substream QoS specifications and priorities.

When a short fading period is observed, a mobile terminal tries to prioritize the transmission of its substreams in order to achieve a minimum QoS. Here, depending on channel conditions, a substream might be dropped for a period of time in order to accommodate higher-priority substreams. To determine the transmission time of any packet in a specific substream (or its position in the transmission queue), the scheduler takes two factors into account: (1) the relative importance of the substream compared to other substreams, and (2) wireless channel conditions.

To achieve both QoS (e.g., bounded delay and reserved bandwidth) and fairness, algorithms like packet fair queueing have to be employed [4]. While the existing packet fair queueing algorithms provide both bounded delay and fairness in wired networks, they cannot be applied directly to wireless networks. The key difficulty is that in wireless networks sessions can experience location-dependent channel errors. This may lead to situations in which a session receives significantly less service than it is supposed to receive, while another receives more. This results in large discrepancies between the sessions' virtual times, making it difficult to provide both delay-guarantees and fairness simultaneously.

To apply packet fair queueing algorithms, Ng et al., [15] identified a set of properties, called Channel-condition Independent Fair (CIF), that a packet fair queueing algorithm should have in a wireless environment: (1) delay and throughput guarantees for error-free sessions, (2) long term fairness for error sessions, (3) short term fairness for error-free sessions, and (4) graceful degradation for sessions that have received excess service. Then they presented a methodology for adapting packet fair queueing algorithms for wireless networks and applied the methodology to derive an algorithm based on the start-time fair queueing [8], called Channel-condition Independent packet Fair Queueing (CIF-Q), that achieves all the above properties [15].

As as an example, we consider two-layer video. Suppose that a subcontract for the base layer specifies reserved bandwidth while a subcontract for the enhancement layer does not specify any QoS guarantee, which is a typical case. We design an architecture for substream scheduling shown in Fig. 3.

Under our architecture, we partition the buffer pool (i.e., data memory in Fig. 3) into two parts: one for base-layer substreams, and one for enhancement layer substreams. Within the same buffer partition for base or enhancement layer, we employ per flow queueing for each substream. Furthermore, substreams within the same buffer partition share the buffer pool of that partition while there is no buffer sharing across partitions. We believe this approach offers an excellent balance between traffic isolation and buffer sharing.

Under the above buffering architecture, we design our per-flow based traffic management algorithms with the aim of achieving requested QoS and fairness. The first part of our architecture is CAC and bandwidth allocation. Video connections are admitted by CAC based on their base-layer QoS requirements. And bandwidth reservations for the admitted base-layer substreams are made accordingly. For admitted enhancement layer substreams, their bandwidth will be dynamically allocated by a bandwidth manager, which has been addressed in Section 3.4. The scaled enhancement layer substreams enter a shared buffer and are scheduled by a First-In-First-Out (FIFO) scheduler. The second part of our architecture is packet scheduling. Shown in Fig. 3 is a hierarchical packet scheduling architecture where a priority link scheduler is shared among a CIF-Q scheduler for base-layer substreams, and an FIFO scheduler for enhancement layer substreams. Service priority is first given to the CIF-Q scheduler and then to the FIFO scheduler.

3.6 Link-layer error control

To provide quality video over wireless, link-layer error control is required. Basically, there are two kinds of error control mechanisms, namely, forward error correction (FEC) and automatic repeat request (ARQ). The disadvantage of FEC is that FEC is not adaptive to varying channel condition and it

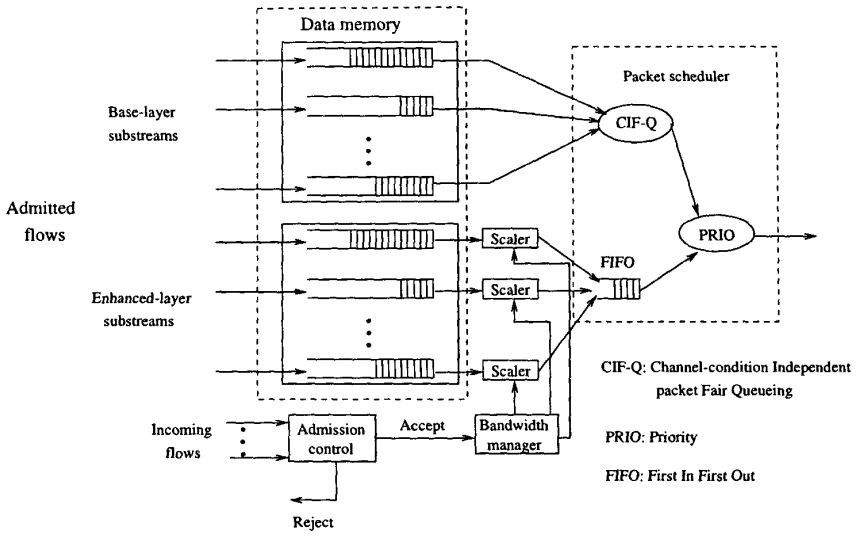

Figure 3: An architecture for substream scheduling at a base station.

works only when BER is below the FEC code's recovery capability. The disadvantage of ARQ is unbounded delay; that is, in the worst case, a packet may be retransmitted in unlimited number of times to recover bit errors.

To address the problems associated with FEC and ARQ, truncated type-II hybrid ARQ schemes [10, 17] and delay-constrained hybrid ARQ [16] have been proposed. These hybrid ARQ schemes combine the good features of FEC and ARQ: bounded delay and adaptiveness. On the other hand, unequal error protection [9] naturally fit the hierarchical structure of scalable video. Specifically, the base layer can be better protected against transmission errors than the enhancement layers. This form of unequal error protection is much more desirable than having to protect all the substreams. An open issue is how to combine unequal error protection with the hybrid ARQ schemes.

4 Summary

Recent years have witnessed a rapid growth of research and development to provide mobile users with video communication through wireless media. In this paper, we examined the challenges in QoS provisioning for wireless video transport. We presented an adaptive framework to support quality video communication over wireless IP networks. The adaptive framework is a combination of network-aware applications and application-aware networks.

The proposed adaptive framework consists of (1) scalable video representations, (2) network-aware video applications, and (3) adaptive services. Under this framework, the mobile terminal and network elements can adapt the video streams according to the channel conditions and transport the adapted video streams to receivers with acceptable perceptual quality. The advantages of deploying such an adaptive framework are that it can achieve suitable QoS for video over wireless, bandwidth efficiency, and fairness in resource sharing.

The contributions of this paper are (1) an adaptive framework including three components, especially, network-aware applications, and (2) a scheduling architecture for scalable video streams, which achieves both QoS (bounded delay and throughput guarantee) and fairness.

Our future work will focus on evaluation of the proposed adaptive framework and scheduling architecture. Under the adaptive framework, there are many issues that need to be addressed for implementation purpose. We list some of them as follows.

- We must consider the particular multiple access control protocol (e.g., CDMA or TDMA), modulation, channel allocation and mobile terminal being used [1].
- We also need to take into account how to adapt the rate at link and physical layers [14]. In addition, channel quality feedback mechanisms have been defined in link/physical layer stan-

dards to carry out rate adaptation. As of the emerging broadband wireless networks, we might also need to design new rate adaptation techniques.

- A scalable video coding scheme has to be carefully designed so that it is robust to multiple time-scale QoS fluctuations in the wireless/wireline network [6]. A scalable video coding scheme should achieve high efficiency with less complexity. It should try to optimally decompose video into multiple substreams without loss of compression efficiency.

As a final note, the adaptive framework is targeted at quality video transport over near-term QoS-enabled wireless IP networks. In addition, the adaptive service could be provisioned either at a single base station or for the entire network. In the real interconnected wireless IP networks, even though we cannot require each router deploy the adaptive service, a partial deployment of the adaptive service can still have clear benefits. For example, a service provider can deploy the adaptive service in its own network and its customers can enjoy the quality offered by the adaptive service in this network. Furthermore, it is entirely feasible to fully deploy the adaptive service within a single administrative domain (e.g., Intranet) and achieve high statistical multiplexing gain and acceptable QoS.

References

[1] I.F. Akyildiz, J. McNair, L.C. Martorell, R. Puigjaner, and Y. Yesha, "Medium access control protocols for multimedia traffic in wireless networks," *IEEE Network Mag.*, pp. 39–47, July 1999.

[2] A. Balachandran, A.T. Campbell, and M.E. Kounavis, "Active filters: delivering scalable media to mobile devices," *Proc. Seventh International Workshop on Network and Operating System Support for Digital Audio and Video (NOSSDAV'97)*, May 1997.

[3] K. Balachandran, S. Kadaba, and S. Nanda, "Rate adaptation over mobile radio channels using channel quality information," *Proc. IEEE GLOBECOM'98*, Nov. 1998.

[4] V. Bharghavan, S. Lu, and T. Nandagopal, "Fair queuing in wireless networks: issues and approaches," *IEEE Personal Commun. Mag.*, pp. 44–53, Feb. 1999.

[5] G. Bianchi, A.T. Campbell, and R. Liao, "On utility-fair adaptive services in wireless networks," *6th International Workshop on Quality of Service (IWQOS'98)*, Napa Valley, CA, May 1998.

[6] Y.-C. Chang and D.G. Messerschmitt, "Adaptive layered video coding for multi-time scale bandwidth fluctuations," submitted to *IEEE J. on Selected Areas in Communications*.

[7] T. Ebrahimi and M. Kunt, "Visual data compression for multimedia applications," *Proceedings of the IEEE*, vol. 86, no. 6, pp. 1109–1125, June 1998.

[8] P. Goyal, H. M. Vin, and H. Chen, "Start-time fair queuing: a scheduling algorithm for integrated service access," *Proc. ACM SIGCOMM'96*, Aug. 1996.

[9] J. Hagenauer and T. Stockhammer, "Channel coding and transmission aspects for wireless multimedia," *Proceedings of the IEEE*, vol. 87, no. 10, pp. 1764–1777, Oct. 1999.

[10] H. Liu and M. El Zarki, "Performance of H.263 video transmission over wireless channels using hybrid ARQ," *IEEE J. on Selected Areas in Communications*, vol. 15, no. 9, pp. 1775–1786, Dec. 1997.

[11] S. Lu, K.-W. Lee and V. Bharghavan, "Adaptive service in mobile computing environments," *5th International Workshop on Quality of Service (IWQOS'97)*, May 1997.

[12] S. McCanne, V. Jacobson, and M. Vetterli, "Receiver-driven layered multicast," *Proc. ACM SIGCOMM'96*, pp. 117–130, Aug. 1996.

[13] M. Naghshineh and M. Willebeek-LeMair, "End-to-end QoS provisioning in multimedia wireless/mobile networks using an adaptive framework," *IEEE Communications Magazine*, pp. 72–81, Nov. 1997.

[14] S. Nanda, K. Balachandran, and S. Kumar, "Adaptation techniques in wireless packet data services," *IEEE Communications Magazine*, pp. 54–64, Jan. 2000.

[15] T.S.E. Ng, I. Stoica and H. Zhang, "Packet fair queueing algorithms for wireless networks with location-dependent errors," *Proc. IEEE INFOCOM'98*, pp. 1103–1111, March 1998.

[16] D. Wu, Y.T. Hou, Y.-Q. Zhang, W. Zhu, and H.J. Chao, "Adaptive QoS control for MPEG-4 video communication over wireless channels," *Proc. IEEE ISCAS'2000*, Geneva, Switzerland, May 28–31, 2000.

[17] Q. Zhang and S. A. Kassam, "Hybrid ARQ with selective combining for fading channels," *IEEE J. on Selected Areas in Communications*, vol. 17, no. 5, pp. 867–880, May 1999.