Scheduling and Adaptive Transmission for the Downlink in 4G Systems

Nilo Casimiro Ericsson[1], Sorour Falahati[2], Anders Ahlén[1], and Arne Svensson[2]

1: Signals and Systems, Department of Materials Science, Uppsala University, SE-751 20 Uppsala, Sweden
E-mail: {nce, anders.ahlen}@signal.uu.se
2: Communication Systems Group, Department of Signals and Systems, Chalmers University of Technology, SE-412 96 Göteborg, Sweden
E-mail: {sorour.falahati, arne.svensson}@s2.chalmers.se

Abstract

In this paper, we propose a system suitable for efficient packet data transmission using the Internet Protocol (IP) in wireless 4G systems. It consists of a hybrid type-II Automatic Repeat reQuest (ARQ) scheme combined with an Adaptive Modulation System (AMS) and a time slot scheduler supplied by channel predictions. It is referred to as predictive HARQ-II/AMS. The performance of the system in a fast fading downlink channel is investigated through simulations where in particular, BER and throughput performance are studied. The proposed system is compared to two simpler systems. One has access to channel predictions and uses adaptive modulation but does not employ error correction and retransmission at the link layer (referred to as predictive AMS). The other has no access to the channel prediction but use HARQ-II/AMS in a blind way (referred to as blind HARQ-II/AMS). The results show that with perfect channel prediction both predictive AMS and predictive HARQ-II/AMS satisfy the BER requirements with some advantages to the former in terms of efficient usage of channel capacity. However, with imperfect channel prediction a retransmission protocol is required to meet the BER requirement making predictive HARQ-II/AMS the winner.

1 Introduction

Packet based traffic over wireless links, using IP is a major concern for future 4G communications. In general, TCP/IP is designed for a highly reliable transmission medium in wired networks where packet losses are rare and are interpreted as congestion in the network. On the contrary, a wireless network employs time varying channels where packet losses may be common due to severe fading. This is misinterpreted by TCP as congestion which leads to inefficient utilization of the available radio link capacity. The result is significant degradation of the wireless system performance [1–3].

Exploiting efficient algorithms at lower layers of wireless networks such as retransmission protocols (ARQ), adaptive error correction, and adaptive modulation combined with channel prediction can significantly improve the communication reliability and throughput. In this way, the lower layers are adapted to the channel conditions while still providing some robustness through retransmission. Moreover, the system performance can be further improved by utilizing an efficient time slot scheduler which shares the spectrum efficiently between users while satisfying the required QoS requirements [4–8].

In this work we propose a HARQ-II/AMS combined with a time slot scheduler supplied by channel predictions. Basically, we assume that the channel quality for each radio link can be predicted for a time interval of about 10 ms into the future [9,10], and that these predictions are accessible by the link layer. Based on the predicted values, the HARQ-II/AMS preliminary selects a Modulation and Coding Scheme (MCS) for each user which satisfies the BER requirement and provides high throughput. The scheduler uses the information about the individual data streams, along with the predicted values of the different radio links and selected MCSs by the link layer to distribute the time slots among the users. The planning is performed so that the desired QoS and priority associated to different users are guaranteed while the channel spectrum is efficiently utilized.

In Section 2 the proposed system is described in detail. The simulation assumptions are explained in Section 3 where the results are presented and discussed. Finally, some conclusions are drawn in Section 4.

2 System description

To evaluate our approach we have chosen to mix several types of traffic. They all constitute sessions that exist for a predefined amount of time. Each session can have either deterministic or random values for packet sizes and packet inter-arrivals. When random values are considered, the corresponding distributions must be specified. For our purposes we have defined three traffic classes: *voice*, *data*, and *media*. The traffic is generated using a Poisson distribution for the packet inter-arrival time and a Pareto distributed packet size, except for *voice*, which is chosen to have a fixed packet size [11]. In the following, three major parts of the system referred to as *buffer*, *scheduler* and *hybrid type-II ARQ/AMS*, are described in Sections A, B and C, respectively.

A Buffer

From the wired network we maintain a buffer with separate queues for the different traffic flows. The packets are distinguishing with respect to the destination, so that each source-destination pair has a separate incoming queue. All incoming packets are scanned for size, priority, session identification number, and required service level from the link layer.

This information is stored in the buffer and can be accessed by the scheduler. The buffer controller is assumed to be able to submit a status report to the scheduling subsystem, described in Section B, so that the scheduler can make an appropriate decision on which queues to choose for the next transmission frame.

The queues are emptied in a bit-by-bit manner, independently of the individual packet boundaries. The bit-stream is passed to the link layer, along with information about the service requirements. The incoming buffer is visualized in Fig. 1. At the receiving side of the wireless link, the packets have to

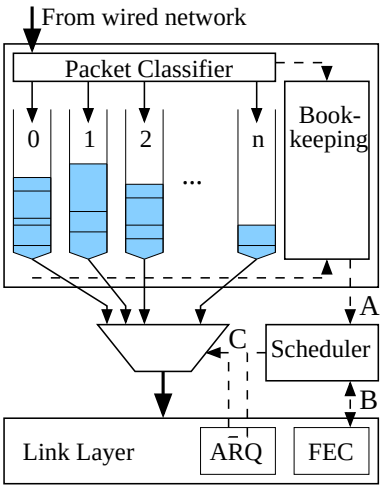

Figure 1: Schematic of the buffer and its queues, and how they interconnect to the scheduler and link layer.

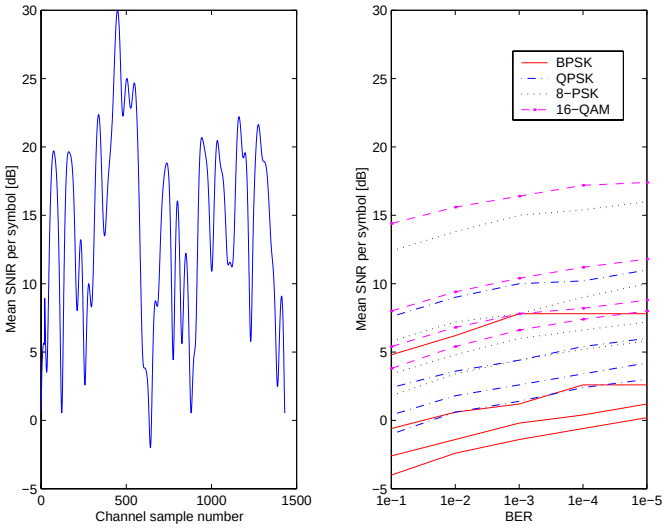

Figure 2: SNR trend and the bit error probability for different combinations of modulation and coding rates. For a fixed modulation, the BER corresponding to the different codes is decreasing by decreasing the coding rate.

be re-assembled, before passing them up to the network layer. This can be done since the scheduling decision is transmitted (broadcasted) to the receiving side, and it totally determines which byte belongs to which flow.

B Scheduler

The system we propose makes use of a channel predictor and a multi-user time slot scheduler that are mounted on top of a HARQ-II/AMS scheme. The scheduler creates a signaling pipe [3] between the network layer buffer and the link layer service, making them mutually aware of each another. For instance, the network layer does not ask for a link service whenever there is data to transmit. Instead it notifies the scheduler of the incoming traffic by passing queueing information (A in Fig. 1) about the amount of data and type of service that would be preferred by the packets. The scheduler then asks the link layer for a report (B in Fig. 1) about how the channel conditions would meet the required service. This can be done since the link layer has access to channel prediction data of all the established connections.

The link layer builds up an $M \times N$ matrix of channel quality predictions, where M is the number of time slots, and N is the number of traffic streams with ongoing sessions. These predictions cover the following time frame of 5 ms. The predictor has a prediction horizon of 10 ms. Fig. 2 shows the predicted values for one link. Based on the predicted values of the Signal to Noise Ratio (SNR) for each of the N user's channels, and the target error probabilitites, the initial code rate and signaling constellation are chosen in advance by the link layer, for a set of $M = 48$ future time slots[1]. It is then the task for the scheduler to, along with the queue sizes and priorities, distribute these time slots among the different queues in a fashion that maximizes some criterion, such as throughput or another measure of user satisfaction [11].

The scheduler performs the scheduling in two rounds. In the first round, each time slot is simply allocated to the user that can transmit at the highest rate in that time slot. If the buffered data in the queues had infinite size, this approach would actually maximize system throughput. However, since the buffers are not infinite, the scheduler runs a second round, where time slots are redistributed from users that have been over-supplied (rich), to users that have been under-supplied (poor). We call this equalization to user satisfaction the *Robin Hood* principle: *"to take from the rich, and give to the poor"*.

C Hybrid type-II ARQ/AMS

Even though the intention is to achieve reliable communication through suitable allocation of time slots by the scheduler as well as appropriate selection of MCS by the physical layer, a given BER can not be guaranteed due to prediction errors. Therefore, a selective repeat ARQ protocol at the link layer needs to cooperate with the adaptive MCS at the physical layer. We refer to this scheme as hybrid type-II ARQ/AMS (HARQ-II/AMS). The ARQ protocol has a limited number of retransmissions to fulfil delay requirements. Further error recovery and complete end-to-end reliability is accomplished by the error and flow control mechanism defined at TCP which is not within the scope of this work.

HARQ-II/AMS is based on a family of Rate Compatible Convolutional (RCC) codes with a parent code of rate 1/3 and memory 6. The higher rate codes are optimum punctured codes while the lower rate codes are optimum repetition codes [12]. The puncturing and repetition matrices have period 2. At each retransmission only incremental redundancy (the coded bits in the lower rate code that are not part of the coded bits in the higher rate code) is transmitted. The transmitter chooses a modulation scheme among the alternatives 16-QAM, 8-PSK, QPSK, and BPSK for each new transmission.

The selection of MCS is based on the user target BER and the predicted SNR value which is available at the link layer. For each MCS, the minimum required SNR values (referred to

[1] In an optimum scheme the transmitted power should also be adjusted to give exactly the desired BER, but this is not done here [5].

as the *SNR limits*) for satisfying different BER requirements are specified, accordingly. For a given target BER, the corresponding *SNR limits* are compared with the predicted SNR, to select the MSC which provides the maximum throughput.

An analytical evaluation of the optimum HARQ-II/AMS and the corresponding *SNR limits* has proven to be difficult due to the the large signalling constellations, the convolutional encoding, and the incremental redundancy transmission. Therefore a suboptimum numerical approach is preferred in this work [12]. The simulated BER is evaluated for all the combinations of 16-QAM, 8-PSK, QPSK, and BPSK modulations and the code rates 1, 2/3, 1/2, 2/5, 1/3, 1/4, 2/9, 1/5, 2/11 and 1/6 for an AWGN channel which is the presumed channel model within each time slot. Some of the simulated *SNR limits* are illustrated in Fig. 2 along with a realization of SNR of a channel. Based on these results an appropriate MCS can be chosen for the first transmission. For the incremental redundancy transmission an optimum selection is more difficult since it depends on what MCSs were used in previous transmissions. In this paper the same *SNIR limits* as in the first transmission have been used also in the retransmissions. To fulfil the rate compatibility criterion, only MCSs with smaller code rates are allowed in retransmission. This choice leads to a somewhat reduced throughput.

The selection of MCS is performed in two phases as described in the following:

Phase one happens at point B in Fig. 1, when the scheduler demands a report from the link layer. This report contains the channel prediction values for all the users with ongoing sessions with corresponding selected MCSs and the temporary priorities (i.e. the users with retransmission request).

Phase two occurs at point C in Fig. 1, when the scheduler informs the link layer about the decision for the time slot allocation, by reporting which users have been allocated time slots and which time slots.

At *Phase two* where the final decision for MCS is performed, two simplifying design assumptions are used. First, a *fixed modulation* is used within a time slot while variable coding rate is allowed for different packets. The second one is the constraint on *rate compatibility* for retransmitting packets. For each user, the highest priority is assigned to retransmission requests (if they exist), following the policy of *"first in, first out"*. These packets are dynamically assigned to the time slots, meaning that among the time slots which can offer an appropriate MCS, the one offering the maximum throughput is selected. In this fashion, the retransmitting packets are occupying different time slots with corresponding MCSs. In case of failure, the transmission of the erroneous packet is postponed to the next frame.

After this stage, based on the partial or completely emptiness of the time slots for each user and the appropriate MCS, the link layer drains *new data* from the corresponding buffers (the arrow at C in Fig. 1). Hence, new packets are formed which fill the corresponding time slots.

After completing this procedure for all the users, the frame signal is generated by Cyclic Redundancy Check (CRC) coding for error detection, convolutional encoding for error correction, and modulation for each individual data packet at the physical layer. More details about the frame structure is given in Section D. At each mobile host, the receiver performs optimum soft decoding by a Viterbi decoder for the parent code where metrics calculations are done according to the puncturing or repetition scheme for the given code rate. The metric calculation is based on Channel State Information (CSI). The decoded bits are fed into the CRC decoder. In case of error detection, a retransmission request (NACK or RQ) is fed back to the link layer transmitter whenever retransmission is permitted, and otherwise an acknowledgement (ACK) is fed back.

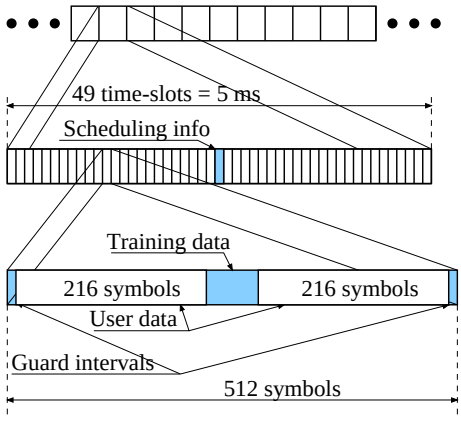

Figure 3: Frame structure used in this paper.

The channel is assumed to be a time-varying flat Rayleigh fading channel. The fading is assumed slow enough for the SNR to be considered constant within each time slot. Between time slots the fading correlation is assumed to fulfil the Jakes correlation model [12]. This corresponds to assuming an AWGN channel in each time slot, but with different SNRs in different time slots.

D Frame structure

Each time slot comprises of user data, and a mid-amble of training data to aid the channel estimation and prediction process. Before and after the user data there are guard intervals to take care of bad timing in the reception or transmission. A descriptive image of the time slot format is given in Fig. 3. This figure also shows the location of the scheduling information within the frame.

A frame consists of 49 time slots, each of which can be dynamically assigned to one user, except for one of the time slots in each frame, which is dedicated to broadcast of scheduling information for the next frame. Furthermore, the assumption of a symbol rate at 5 Msymbol/s which is used in this work, allows each time slot to contain 512 symbols[2]. Please note that a time slot may consist of several packets. A compromise between the base-station and the mobile terminal leads to the suggestion that the scheduling information is transmitted in one of the time slots in the middle of the frame. In this way the scheduler has time to perform the necessary calculations, and

[2]A mobile radio channel with symbol rate of 5 Msymbol/sec is typically frequency selective, but in this paper we assume for simplicity that it is flat.

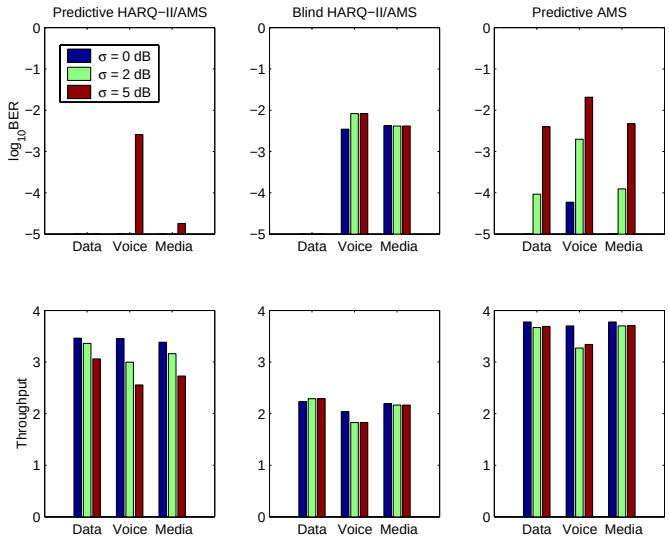

Figure 4: The BER and throughput (bits/symbol) for the *predictive HARQ-II/AMS*, *predictive AMS* and *blind HARQ-II/AMS* for σ, the standard deviation of the prediction error, equal to 0, 2 and 5 dB. The users belong to the classes *voice* (BER=10^{-3}), *data* (BER=10^{-5}), and *media* (BER=10^{-4}).

the mobile stations have time to adjust to the new schedule.

Since the scheduling information that is transmitted in the downlink is crucial for the efficiency of the multiple access scheme, it has to be well protected against errors. For this time slot, only BPSK modulation with a low rate code should be used. In this work we assume that scheduling information is correctly received.

3 Simulation results

The following assumptions are used during the simulation. The data packets at the link layer contain 216 bits, including 12 CRC bits for error detection and 6 zero tail bits corresponding to the memory of the convolutional encoder. 16-QAM, 8-PSK, QPSK, and BPSK modulations are employed where Gray coding is used for mapping the bits to the symbols. The maximum symbol energy and symbol rate are presumed to be constant in all the modulation schemes. The channel SNR prediction is assumed to be lognormally distributed with a mean value equal to the true SNR and a standard deviation of σ dB. The CSI used at the receiver is assumed to be estimated without error. Moreover, the channel is AWGN within a time slot but fading during a frame transmission. For *voice*, *media* and *data* traffic classes, target BERs of 10^{-3}, 10^{-4} and 10^{-5} are presumed, respectively. Additionally, the maximum allowed number of retransmissions at the link layer is chosen to be 3, 3, and 8 for *voice*, *media* and *data* traffic classes, respectively. Each simulation run corresponds to 0.146 s.

For the performance assessment of the proposed system which is referred to as the *predictive HARQ-II/AMS* and in order to investigate the effect of coding, ARQ, and channel prediction at the transmitter, we have considered two other systems for comparison. One is the so-called the *predictive AMS* where no coding and retransmission protocols are allowed at lower layers but the link layer is provided with the channel predictions and an adaptive modulation system, similar to the one presented in [11]. In this system 16-QAM, 8-PSK, QPSK, and BPSK modulations are used where the corresponding *SNR limits* are analytically evaluated as described in [13]. This scheme can be improved by using the techniques given in [5], but this is not pursued further here. The *predictive AMS* reports the channel predictions and suggested modulation schemes to the scheduler for time slot allocation. Furthermore, since no coding is involved here, the data packets at the link layer generally contain 6 extra data bits due to the absence of tail bits, compared to the *predictive HARQ-II/AMS*.

The second scheme for comparison is the so-called *blind* or *non-predictive HARQ-II/AMS* where the link layer does not have access to any channel predictions and selection of the MCS is based on predetermined values stored in a look-up table, similar to the one presented in [12]. The chosen MCS scheme is here independent of the channel. In this system, the transmission starts using a modulation with large signalling constellation and a high rate code. During the retransmission, the size of constellation as well as the coding rate are reduced. For the *data* traffic class, the transmission is initialized with 16-QAM at rate 1. The next transmission attempts are performed by 16-QAM at rate $2/3$, followed by 8-PSK at rates $1/2$ and $2/5$, QPSK at rate $1/3$, $1/4$ and $2/9$ and finally BPSK at rate $1/5$. For *media* and *voice* traffic classes, the transmission starts with 16-QAM at rate 1, followed by 8-PSK at rates $2/3$ and $2/5$. Moreover, only the information according to temporary priorities due to the users with retransmission request is provided to the scheduler.

The simulations are carried out for 15 users with 5 users belonging to each of the traffic classes of *voice*, *data*, and *media*. The BER and throughput performance of three systems is evaluated both for perfect ($\sigma = 0$ dB) and imperfect ($\sigma = 2$ and 5 dB) channel predictions. Here, the throughput is defined as the number of accepted received data bits per transmitted symbol[3]. The throughput and bit error probability (BER) results are averaged over the users within each of the *voice*, *data*, and *media* classes and are given in Fig. 4.

The results show that with perfect channel predictions, both the *predictive HARQ-II/AMS* and *predictive AMS* meet the BER requirements in contrast to the *blind HARQ-II/AMS* with maximum three transmissions, i.e. *voice* and *media* traffic classes[4]. This stems from the fact that having access to the information about the future characteristics of the channel results in a better choice of modulation and (or) coding rate. Further improvements are possible by also adjusting the transmit power in an optimum way [5]. However, comparing the throughput and BER performance leads to the conclusion that on average, *predictive AMS* performs well while *predictive HARQ-II/AMS* has lower throughput and an unnecessary low BER. The *blind HARQ-II/AMS* looses a lot in throughput since it does not utilize channel predictions. The conclusion is that

[3]Please note that in *predictive* or *blind HARQ-II/AMS*, only the packets without any detectable errors are regarded as accepted ones. In *predictive AMS*, all the received packets are considered to be accepted since no error detection codes are utilized. Some of the accepted packets contain bit errors which contribute to the bit error probability.

[4]With a larger number of retransmission like 8 for *data*, the BER requirements are fulfilled also with the *blind HARQ-II/AMS*.

with ideal channel prediction, the required error probability can be obtained without ARQ and channel coding, which is not surprising. In a practical system, perfect channel predictions are however never achievable.

The importance of the ARQ protocol and channel coding appears when erroneous channel predictions are introduced. Now ARQ with a larger number of retransmissions is needed to guarantee a given BER. Since our schemes allow only a relatively small number of retransmissions for *voice* and *data*, none of them really can guarantee a given BER for these services. In the results shown in Fig. 4 however, *predictive HARQ-II/AMS* fulfils the BER requirement except for *voice* with $\sigma = 5$ dB. Throughput is however lost with larger prediction errors. The *blind HARQ-II/AMS* is not affected by the prediction errors, while the *predictive AMS* have a high throughput also with prediction errors, but this is at the expense of a too large BER.

4 Conclusion

A combination of radio channel predictive time slot scheduling and adaptive HARQ-II/AMS for IP packet data on wireless channels is presented and evaluated through simulations. The performance of the proposed system referred to as the *predictive HARQ-II/AMS* is evaluated. Comparison is made with a system where no channel state information is available at the transmitter (the *blind HARQ-II/AMS*) and a system using channel prediction but no channel coding and no repetition protocol (the *predictive AMS*). Three traffic classes referred to as *voice*, *data*, and *media* are chosen where each requires different BER.

The conclusion from this paper is that a repetition protocol must be used with prediction errors, or a guaranteed BER can not be obtained. This however leads to a reduced throughput and a larger delay in the system. With no prediction error, the coding and modulation can be chosen to guarantee the required BER in a single transmission which optimize the throughput and delay [5]. In practice however, a nonzero prediction error will always be present. In such a situation, our scheme can be further improved. One problem is that the adaptive MCS is designed for the ideal case with no prediction error. The throughput can be increased and the delay reduced if the MCS can be designed for a given prediction error model and a given channel estimation error model which are available at the transmitter. This is a topic that we are currently working on.

Another possible improvement is in the scheduling. Our current scheduler is quite adhoc and optimization should be possible. It is however not straight forward to define a suitable optimization criterion but work is in progress on this topic. In this paper we have assumed a flat wideband channel which is not that common. The ideas presented can however quite easily be generalized to frequency selective channels by using Orthogonal Frequency Division Multiplexing (OFDM) techniques [13]. Another assumption is that the system has no frequency reuse and therefore no cochannel interference. This is a serious restriction, since we believe it is much more difficult to predict the signal to noise plus interference ratio (SNIR) in a cochannel environment with intermittent packet transmission on the cochannels. Work on SNIR prediction in this environment will be pursued in the future. The performance and robustness in the presence of slow fading and SNR variations on a longer time horizon also remains to be studied. Added robustness will here be provided by suitably designed transport protocols [14].

The ideas presented in this paper are the basis for a more complete system proposal in a companion paper [15].

References

[1] H. Balakrishnan, S. Seshan, E. Amir, and R.H. Katz, "Improving TCP/IP performance over wireless networks," in *Proc. MOBICOM*, Berkeley, CA, 1995, pp. 2–11.

[2] A.S. Tanenbaum, *Computer Networks*, Prentice Hall International, Upper Saddle River, New Jersey, 1996.

[3] G. Wu, Y. Bai, J. Lai, and A. Ogielski, "Interactions between TCP and RLP in wireless internet," in *IEEE Global Telecom. Conf.*, Rio de Janeiro, Brazil, December 1999, pp. 661–666.

[4] C. Roobol, P. Beming, J. Lundsjö, and M. Johansson, "A proposal for an RLC/MAC protocol for wideband CDMA capable of handling real time and non-real time services," in *Proc. IEEE Veh. Techn. Conf.*, May 1998, pp. 107–111.

[5] S.T. Chung and A.J. Goldsmith, "Degrees of freedom in adaptive modulation: a unified view," in *IEEE Trans. Commun.*, vol. 49, no. 9, pp. 1561–1571, Sep. 2001.

[6] M. Kawagishi, S. Sampei, and N. Morinaga, "A novel reservation TDMA based multiple access scheme using adaptive modulation for multimedia wireless," in *Proc. IEEE Veh. Techn. Conf.*, May 1998, pp. 112–116.

[7] D.J. Goodman, R.A. Valenzuela, K.T. Gayliard, and B. Ramamurthi, "Packet reservation multiple access for local wireless communications," *IEEE Trans. Commun.*, vol. 37, no. 8, pp. 885–890, Aug. 1989.

[8] M. Naijoh, S. Sampei, N. Morinaga, and Y. Kamio, "ARQ schemes with adaptive modulation/TDMA/TDD systems for wireless multimedia communication services," in *Proc. IEEE Int. Symp. Personal, Indoor and Mobile Radio Commun.*, Helsinki, Finland, Sept. 1997, pp. 709–713.

[9] T. Ekman, G. Kubin, M. Sternad, and A. Ahlén, "Quadratic and linear filters for mobile radio channel prediction," in *Proc. IEEE Veh. Techn. Conf.*, Amsterdam, The Netherland, Sept. 1999, pp. 146–150.

[10] T. Ekman, *Prediction of Mobile Radio Channels*, Technical licentiate thesis, Dep. of Signals and Systems, Uppsala University of Technology, Uppsala, Sweden, Dec. 2000.

[11] N. C. Ericsson, "Adaptive modulation and scheduling of IP traffic over fading channels," in *Proc. IEEE Veh. Techn. Conf.*, Amsterdam, the Netherlands, Sept. 1999, pp. 849–853.

[12] S. Falahati and A. Svensson, "Hybrid type-II ARQ schemes with adaptive modulation systems for wireless channels," in *Proc. IEEE Veh. Techn. Conf.*, Amsterdam, The Netherlands, Sept. 1999, pp. 2691–2695.

[13] J. G. Proakis, *Digital Communications, 4/ed*, McGraw-Hill, New York, 2001.

[14] A. Ewerlid, "Reliable communication over wireless links," in *Proc. Nordic Radio Symposium*, Saltsjöbaden, Sweden, Apr. 2001.

[15] T. Ottosson, A. Ahlén, A. Brunstrom, M. Sternad, and A. Svensson "Towards 4G IP-based wireless systems," in *Proc. Future Telecom. Conf.*, Beijing, China, Nov. 2001.