CCIT Report #660 November 2007

Multi-Layer Broadcasting Hybrid-ARQ Strategies for Block Fading Channels

Avi Steiner and Shlomo Shamai (Shitz)

DRAFT

August 31, 2007

Abstract

Conventional hybrid automatic retransmission request (HARQ) is usually used to maximize throughput. However, high throughput is achieved at the expense of high latency. We study a novel broadcasting HARQ strategy. The multi-layer broadcast approach is suitable for the case where transmitter has no channel state information (CSI), which is the case with HARQ schemes as well. The broadcast approach enables the receiver to decode rates, which are matched to every fading gain realization. That is, the higher the fading gain realization, the more layers are reliably decoded. The broadcast approach combined with HARQ enables achieving high throughput with low latency. In a broadcast HARQ scheme every code layer supports HARQ independently. Thus HARQ is applied in every transmission block to undecoded layers only, which highly increases the broadcast approach efficiency. In this paper, both broadcast chase combining (BCC) HARQ and broadcast incremental redundancy (BIR) HARQ are studied in the limit of infinitely many layers, and for finite level coding. Interestingly, with continuous broadcasting the BCC-HARQ is found to closely approximate the BIR-HARQ, while using a sub-optimal broadcasting power distribution.

Index Terms

Multi-layer broadcasting, hybrid-ARQ, incremental redundancy, chase combining.

I. Introduction

An efficient technique for increasing the average throughput is by using retransmissions based on hybrid automatic retransmission request (HARQ). A basic ARQ scheme, known as

This work was supported by the REMON consortium.

the ALOHA protocol, requires retransmission for bad channel conditions (outage), until the channel is sufficiently good to allow for reliable decoding. Its maximal throughput is known as the outage capacity [1]. A more advanced HARQ scheme performs optimal coherent combining of all retransmissions, thus improving the probability of successful decoding. This HARQ scheme is known as the chase combining HARQ (CC-HARQ) [2]. With incremental-redundancy HARQ (IR-HARQ), retransmissions include additional parity bits which allow joint decoding of previously transmitted blocks, and thus reduce the rate of outage events. In all ARQ strategies, there is a fundamental tradeoff of throughput and latency. The more allowed retransmissions, the higher the achievable average throughput, at the expense of higher latencies.

In the absence of transmit channel state information (CSI), and when considering the average throughput or delay as figures of merit, it is beneficial to use the broadcast approach [3]. The broadcast strategy facilitates reliable transmission rates adapted to the actual channel conditions [3], [4]. The multi-layer broadcast approach hinges on the broadcast channel, which was first explored by Cover [5]. In a broadcast channel, a single transmission is directed to a number of receivers, each enjoying possibly different channel conditions, reflected in their received signal to noise ratio (SNR). Thus, the higher the fading gain, the higher is the achievable rate. HARQ schemes combat the same outage problem, by retransmitting additional information allowing joint decoding of multiple blocks. Evidently, the HARQ approaches end up with increased latencies. In this work we consider a layered transmission, where each layer supports an HARQ retransmission scheme, and thus after every block, the transmitter schedules a retransmission consisting of undecoded layers only. Hence, higher throughput efficiency may be achieved compared to the conventional HARQ.

In [6], an information theoretic analysis of an IR-HARQ scheme is presented for the multiple-access Gaussian collision channel. Numerous contributions on IR-HARQ code design schemes using low-density parity check (LDPC) codes may be found, e.g. [7], [8], [9], and with application to IR raptor codes [10]. In [11], a combined LDPC IR-HARQ is suggested for a MIMO V-BLAST scheme, where a single outer code is used in transmission. The demodulation consists of an MMSE estimator, followed by an LDPC decoder. In a sense, this concept resembles the BIR approach, where layered data is obtained from the V-BLAST spatial streams. However, in BIR a separate encoder is used per layer, and the V-BLAST decoding in [11] is sub-optimal as it does not include successive cancelation.

A practical IR scheme with Turbo codes is analyzed in [12], and extended to a collaborative setting in [13]. A practical Turbo based joint source-channel IR coding scheme is presented in [14], where puncturing is decreased adaptively as long as the source coding rate is greater than its entropy, this scheme is referred to as combined incremental and decremental redundancy. A modified CC-HARQ which applies retransmission of sub-packets only is studied in [15]. The motivation there is to compensate for burst errors, by retransmitting only nearly erased bits. In [16], a CC-HARQ scheme is investigated, where diversity is obtained by changing the bit-to-symbol mapping every retransmission. This allows using the same code and a bit-interleaved coded modulation (BICM) interleaver for all retransmissions.

For the asymptotic case of high SNR, it is of interest to analyze the diversity-multiplexing tradeoff. An additional dimension of delay in IR-HARQ schemes is studied in [17], where the three dimensional diversity-multiplexing-delay tradeoff is fully characterized for a MIMO channel. An achievable region for this tradeoff in presence of relays is presented in [18].

In this work a new broadcast HARQ concept is suggested, where multi-layered coded data is transmitted. A feedback channel from receiver to transmitter simply indicates the highest successfully decoded layer. The system performs CC-HARQ or IR-HARQ for every layer separately. For finite level IR coding, expressions for average throughput are derived. Expressions for the throughput are explicitly obtained for an outage scheme (classical single level coding with IR-HARQ).

The case of many coded layers is studied through the continuous broadcast approach [4]. This approach achieves its highest efficiency already with a single retransmission, when the channel fading gain remains fixed during retransmissions. That is since the receiver feedback indicating the highest decoded layer, also implicitly designates the channel fading gain. Thus the transmitter can adapt the retransmission rate to guarantee zero outage. Several continuous BIR and BCC HARQ protocols are considered, and average achievable rates are derived. Numerical results show significant gains of continuous broadcasting with one retransmission over finite level coding. One of the main advantages of combined layering and HARQ is the high efficiency achieved with small delays. It is well known that high throughput of conventional HARQ is achieved when many retransmissions are allowed, which requires high average delays. Here, the more transmitted layers, the better the transmitter can reschedule retransmission to provide nearly zero outage.

The rest of the paper is organized as follows. The channel model is presented in section II. Finite level coding HARQ is studied in section III. Continuous BCC-HARQ and BIR-HARQ schemes are derived in IV. Numerical results demonstrating the efficiency of the various protocols are presented in V. Finally, section VI concludes the paper.

II. Channel Model

Consider the following single-input single-output (SISO) channel,

$$\mathbf{y} = h\mathbf{x} + \mathbf{n}, \tag{1}$$

where **y** is a received vector (boldfaced letters are used for vectors). **x** is the original source transmitted vector. **n** is the additive noise vector, with elements that are complex Gaussian i.i.d with zero mean and unit variance, denoted $\mathcal{CN}(0,1)$, and h is the (scalar) fading coefficient. The realization of h is assumed to be perfectly known by the receivers. It remains fixed over a transmission block, and over multiple blocks corresponds to the complex Gaussian distribution, denoted $h \sim \mathcal{CN}(0,1)$. The source transmitter has no CSI, and the power constraint at the source is given by $E|x|^2 \leq P$, where E stands for the expectation operator. Two main channel models are considered:

1) **Long-term static channel (LTSC)** - during all HARQ retransmissions h remains fixed. This model represents a slowly fading channel with low delay HARQ mechanism.
2) **Short-term static channel (STSC)** - in every HARQ retransmission h changes according to the i.i.d $h \sim \mathcal{CN}(0,1)$. This model corresponds to a relatively rapidly varying channel, or the slow fading case with delayed HARQ retransmissions.

III. Finite Level Coding

In this section the achievable throughput of finite level coding combined with IR coding is derived. We begin with a short overview of single level coding (outage) IR-HARQ. Then, a two level coding IR-HARQ is studied.

A. Outage Approach

In an outage approach a single level coding scheme is used in transmission of the first block. If the receiver decodes the first block successfully, it returns a positive acknowledge (ACK);

otherwise it returns a negative acknowledge (NACK). In an IR-HARQ scheme, every time a NACK is received at the transmitter, it schedules a retransmission of the same data with different parity bits, allowing the receiver to decode the original message using jointly all retransmissions. Usually, a maximal number of retransmissions M is defined. If a NACK is returned after M retransmissions, then an outage event is declared.

The average throughput for an outage approach with IR coding follows from [6]. Define a code rate R_1, and assume that the rate on the first block is MR_1 Nats per channel use. Following from the renewal theorem [20], the average throughput is given by

$$\eta = E[\mathcal{R}]/E[D] \tag{2}$$

where $E[\mathcal{R}]$ is the average reward, and $E[D]$ is the expected inter-renewal time. In an outage approach with no HARQ the transmitter sends M blocks of data for every packet, and the receiver attempts decoding only after receiving all M blocks. On decoding failure a NACK is returned, which indicates an outage event. Thus in an outage approach the average reward is

$$E[\mathcal{R}] = MR_1(1 - p_{o,1}(M)) \tag{3}$$

where $p_{o,1}(m)$ is the outage probability on m^{th} retransmission. The delay per transmission is clearly $E[D] = M$. Thus, the average throughput of a conventional outage approach is

$$\eta_1(M) = R_1(1 - p_{o,1}(M)). \tag{4}$$

For an outage approach IR-HARQ the average reward is also (3). However, since the receiver attempts decoding after every received block, the decoding latency may be shorter in case of successful decoding before the M^{th} block. The average inter-renewal delay (latency) is

$$E[D] = \sum_{m=1}^{M} m \cdot q_1(m) + Mp_{o,1}(M) = 1 + \sum_{m=1}^{M-1} p_{o,1}(m) \tag{5}$$

where $q_1(m)$ is defined as the probability of success on the m^{th} retransmission (while failing all $m-1$ previous retransmissions). The second equality in (5) may be verified by noticing that $q_1(m) = p_{o,1}(m-1) - p_{o,1}(m)$, see also [6]. In an IR coding scheme the m^{th} outage probability is

$$\begin{aligned} p_{o,1}(m) &= \Pr\left(I(x;y|h_1) < MR_1, I(x;y|h_1) + I(x;y|h_2) < MR_1, ..., \sum_{k=1}^{m} I(x;y|h_k) < MR_1 \right) \\ &= \Pr\left(\sum_{k=1}^{m} I(x;y|h_k) < MR_1 \right) \end{aligned} \tag{6}$$

where the second equality is due to the monotonicity of the cumulative mutual information. The mutual information $I(x;y|h)$ is specified by

$$I(x;y|h) = \log(1+|h|^2 P) \equiv \log(1+sP) \tag{7}$$

where $s = |h|^2$, and P is the average transmission power, which is also the SNR as in (1).

We consider first the LTSC model, which is defined in section II. In this case $h_k = h$, $\forall k = 1,..M$. Hence, the outage probability on the m^{th} block is given by

$$p_{o,1}(m) \quad = \Pr\left(m \log(1+sP) < MR_1\right) = \Pr\left(s < \frac{e^{MR_1/m}-1}{P}\right) = 1-\exp\left\{-\frac{e^{MR_1/m}-1}{P}\right\} \tag{8}$$

where the last equality subsumes a Rayleigh fading channel, i.e. the fading power cumulative distribution function (cdf) is given by $F(x) = 1 - e^{-x}$, for $x \geq 0$. The average throughput is

$$\eta_{1,IR,LTSC}(M) = \frac{MR_1 \exp\left\{-\frac{e^{R_1}-1}{P}\right\}}{M - \sum\limits_{m=1}^{M-1} \exp\left\{-\frac{e^{MR_1/m}-1}{P}\right\}} \tag{9}$$

where $\eta_{1,IR,LTSC}(M)$ as $\eta_1(M)$ in (4) is the average achievable throughput for the classical IR-HARQ scheme over a Rayleigh fading channel, under the LTSC model.

Using the STSC model, which is defined in section II, the outage probability of an IR-HARQ scheme for $M > 1$ does not lend itself to a closed form solution. Therefore, only the case of $M = 2$ is considered. The case of $M = 2$ is particularly interesting when considering broadcasting over an IR-HARQ scheme, as will be elaborated in section IV. Under the STSC model

$$\begin{aligned} p_{o,1}(2) \quad &= \Pr\left(\log(1+s_1 P) + \log(1+s_2 P) < 2R_1\right) \\ &= \int\limits_0^{(e^{2R_1}-1)/P} ds \ \left(1-\exp\left\{-(\tfrac{e^{2R_1}}{1+sP}-1)/P\right\}\right) e^{-s}. \end{aligned} \tag{10}$$

The average throughput is then specified by $\eta_{1,IR,STSC}(2) = \frac{2R_1(1-p_{o,1}(2))}{1+p_{o,1}(1)}$.

B. Two Level Coding

In what follows we consider multi-layer coding combined with an IR-HARQ scheme. The transmitter performs multi-layer coding for the first transmission. Then, if all layers were reliably decoded, it returns a simple ACK. However, if only some of the layers were decoded (due to the instantaneous fading), it returns a NACK with an index pointing to the highest decoded layer. Thus, the retransmission consists of additional parity bits **only for undecoded layers**.

Our system model assumes a short term power constraint, i.e. identical power constraint P per transmission. This means for layered IR that in retransmission of additional parity bits for undecoded layers, the power associated with a specific layer increases every retransmission as more and more layers are decoded. In case of a long term average power constraint, which is defined on multiple transmission blocks, the throughput may be even further optimized by using a different power allocation for each retransmission. However, such power allocation is not in the scope of this work.

We derive now the expected throughput, as defined in (2), for two level coding HARQ. The average reward is

$$E[\mathcal{R}] = MR_1(1 - p_{o,1}(M)) + MR_2(1 - p_{o,2}(M)) \tag{11}$$

where $p_{o,i}(m)$ is the outage probability on the m^{th} retransmission of the i^{th} layer. The average inter-renewal delay is

$$E[D] = \sum_{m=1}^{M} m \cdot q_2(m) + Mp_{o,2}(M) \tag{12}$$

where $q_2(m)$ is defined as the probability of successful decoding of the second layer on the m^{th} retransmission (while failing all $m-1$ previous retransmissions).

*1) **LTSC model**:* In this channel model a high (or low) SNR approximation is required for evaluation of $p_{o,2}(m)$, and $q_2(m)$. The outage probability of the first layer is straightforward to derive

$$p_{o,1}(m) \quad = \Pr\left(mI(x_1;y|h) < MR_1\right) = \Pr\left(m\log(1+\tfrac{\alpha sP}{1+\overline{\alpha}sP}) < MR_1\right) = 1 - e^{-\beta(m)} \tag{13}$$

where $\beta(m) \triangleq \frac{e^{MR_1/m}-1}{P\left(\alpha - \overline{\alpha}(e^{MR_1/m}-1)\right)}$, and where αP is the power allocated to the first layer, and $\overline{\alpha}P = (1-\alpha)P$ is the power allocated to the second layer ($0 \leq \alpha \leq 1$) in layered transmissions. In computation of $p_{o,2}(m)$ the mutual information of the second layer depends on whether or not the first layer was reliably decoded. This is formulated in the following probabilistic expression

$$\begin{aligned} p_{o,2}(m) = \sum_{k=1}^{m} \quad & \Pr\{kI(x_1;y|h) \geq MR_1,\ (k-1)I(x_1;y|h) < MR_1, \\ & kI(x_2;y|h,x_1,P_2=\overline{\alpha}P) + (m-k)I(x_2;y|h,x_1,P_2=P) < MR_2\} \\ & +\Pr\left(mI(x_1;y|h) < MR_1\right) \end{aligned} \tag{14}$$

where P_2 is the allocated power to the second layer. As may be noticed for every k, there are $(m-k)$ transmissions of the second layer only, as the first layer is successfully decoded on the k^{th} retransmission. The expression of $p_{o,2}(m)$ in (14) may be simplified as follows

$$\begin{aligned} p_{o,2}(m) &= \sum_{k=1}^{m} \text{Pr}\left(\beta(k) \leq s < \beta(k-1),\ (1+\overline{\alpha}sP)^{k}(1+sP)^{m-k} < e^{MR_2}\right) + p_{o,1}(m) \\ &\approx \sum_{k=1}^{m} \text{Pr}\left(\beta(k) \leq s < \beta(k-1),\ \overline{\alpha}^{k}(sP)^{m} < e^{MR_2}\right) + p_{o,1}(m) \\ &= 1 - e^{-\beta(m)} + \sum_{k=1}^{m} e^{-\zeta_2(m,k)} - e^{-\zeta_1(m,k)} \end{aligned} \tag{15}$$

where $\beta(0)=\infty$, and $\zeta_1(m,k) \triangleq \min\{\beta(k-1), \gamma(m,k)\}$, and $\zeta_2(m,k) \triangleq \min\{\beta(k), \gamma(m,k)\}$, and $\gamma(m,k) \triangleq \frac{1}{P}\overline{\alpha}^{-k/m}e^{MR_2/m}$. In addition, the above approximation holds for $\overline{\alpha}sP >> 1$. The low SNR approximation for $\gamma(m,k)$ is given for $\overline{\alpha}sP << 1$, and is $\gamma(m,k) \triangleq \frac{e^{MR_2}-1}{m(1+\overline{\alpha})P}$. However, in the numerical results we will be using only the high SNR approximation. Similarly, the probability $q_2(m)$ of successful decoding of the second layer at the m^{th} retransmission is

$$\begin{aligned} q_2(m) = \textstyle\sum_{k=1}^{m} \text{Pr}(\quad & \beta(k) \leq s < \beta(k-1), \\ & kI(x_2;y|h,x_1,P_2=\overline{\alpha}P) + (m-k-1)I(x_2;y|h,x_1,P_2=P) < MR_2, \\ & kI(x_2;y|h,x_1,P_2=\overline{\alpha}P) + (m-k)I(x_2;y|h,x_1,P_2=P) \geq MR_2), \end{aligned} \tag{16}$$

which simplifies into

$$q_2(m) = \sum_{k=1}^{m} \text{Pr}(\upsilon_2(m,k) \leq s < \upsilon_1(m,k)) = e^{-\upsilon_2(m,k)} - e^{-\upsilon_1(m,k)} \tag{17}$$

where $\upsilon_1(m,k) \triangleq \min\{\beta(k-1), \gamma(m-1,k)\}$, and $\upsilon_2(m,k) \triangleq \min\{\upsilon_1(m,k), \max\{\beta(k), \gamma(m,k)\}\}$. Notice that the relationship between $p_{o,1}(m)$ and $q_1(m)$, which was established for the single level coding in (5), does not hold for $p_{o,2}(m)$ and $q_2(m)$ in two level coding, as indicated by (15) and (16).

*2) **STSC model:*** Analytical derivation of the probabilities $q_2(m)$, $p_{o,1}(m)$, and $p_{o,2}(m)$ could not be done in closed form, for every M. However, for $M=2$ these probabilities have a single integral form, and are derived as follows. From (11)-(12), it is clear that derivation of $p_{o,1}(2)$ and $p_{o,2}(2)$ is required. We note that $E[D]$, for $M=2$, is $E[D] = 1 + p_{o,2}(1)$, which follows from (12). The outage probability of the second layer on the first transmission is

$$\begin{aligned} p_{o,2}(1) &= \text{Pr}\{I(x_1;y|h_1) < 2R_1 \text{ or } (I(x_1;y|h_1) \geq 2R_1, I(x_2;y|h_1,x_1,P_2=\bar{\alpha}P) < 2R_2)\} \\ &= 1 - e^{-\gamma_1} \end{aligned} \tag{18}$$

where $\gamma_1 = \max\left(\frac{e^{2R_1}-1}{\alpha P - \bar{\alpha}(e^{2R_1}-1)P}, \frac{e^{2R_2}-1}{\bar{\alpha}P}\right)$. The first layer outage probability, for $m = 2$, is

$$\begin{aligned} p_{o,1}(2) &= \mathrm{Pr}\{I(x_1;y|h_1) + I(x_1;y|h_2) < 2R_1\} \\ &= \int\limits_0^\infty ds_2 e^{-s_2}(1 - \exp\left\{-\max\left(0, \frac{\gamma_2}{\alpha P - \bar{\alpha}P\gamma_2}\right)\right\}) \end{aligned} \tag{19}$$

where $\gamma_2 = \frac{e^{2R_1}}{1+\alpha s_2 P/(1+\bar{\alpha}s_2 P)} - 1$. The outage probability of the second layer for $m = 2$, while reliably decoding the first layer, is as follows

$$\begin{aligned} p_{o,2}(2) &= \mathrm{Pr}\{I(x_1;y|h_1) + I(x_1;y|h_2) < 2R_1\} \\ &\quad +\mathrm{Pr}\{I(x_1;y|h_1) < 2R_1 I(x_1;y|h_1) + I(x_1;y|h_2) \geq 2R_1, \\ &\quad I(x_2;y|h_1,x_1,P_2 = \bar{\alpha}P) + I(x_2;y|h_2,x_1,P_2 = \bar{\alpha}P) < 2R_2\} \\ &\quad +\ \mathrm{Pr}\{I(x_1;y|h_1) \geq 2R_1 I(x_2;y|h_1,x_1,P_2 = \bar{\alpha}P) < 2R_2, \\ &\quad I(x_2;y|h_1,x_1,P_2 = \bar{\alpha}P) + I(x_2;y|h_2,x_1,P_2 = P) < 2R_2\} \\ &= p_{o,1}(2) \\ &\quad + \int\limits_0^\infty ds_2 e^{-s_2}(\exp\left\{-\max\left(0, \frac{\gamma_2}{\alpha P - \bar{\alpha}P\gamma_2}\right)\right\} - \exp\left\{-\min\left((\frac{e^{2R_2}}{1+\bar{\alpha}Ps_2} - 1)/\bar{\alpha}P, \gamma_3\right)\right\}) \\ &\quad + \int\limits_0^\infty ds_2 e^{-s_2}(\exp\{-\max(0,\gamma_3)\} - \exp\left\{-\min\left((e^{2R_2}-1)/\bar{\alpha}P, (\frac{e^{2R_2}}{1+Ps_2} - 1)/\bar{\alpha}P\right)\right\}) \end{aligned} \tag{20}$$

where $\gamma_3 = \frac{e^{2R_1}-1}{\alpha P - \bar{\alpha}P(e^{2R_1}-1)}$. The second term in (20) is the probability of decoding the first layer and failing on decoding of the second layer on first transmission. The last term is the probability of decoding the first layer on the first transmission, and failing to decode the second layer after the retransmission, where the retransmission consists of the second layer only with power $P_2 = P$.

The average throughput can now be computed, from (18)-(20) for two level coding, by using $\eta_{1,IR,STSC}(2) = \frac{2R_1(1-p_{o,1}(2)) + 2R_2(1-p_{o,2}(2))}{1+p_{o,2}(1)}$.

IV. A Continuous Broadcast and HARQ Approaches

In this section layered CC-HARQ and IR-HARQ are studied in the limit of continuous layering, such that after the first layered transmission, the transmitter retransmits the same data or additional parity bits for the undecoded layers. The combined layering and HARQ may be practically used with superposition coding schemes, or dirty paper coding schemes. Only LTSC model is considered. Interestingly, under this model and continuous broadcasting, the feedback of the first transmission actually reveals the exact fading gain (within a specific range). This allows the transmitter to retransmit only part of the layered information such that **no outage** will occur on retransmission. Clearly, the range of fading gains revealed to transmitter depends on the power distribution of first transmission. Another important aspect of this approach is that the first retransmission can already guaranty (to some extent) zero outage, which encourages to limit M to $M = 2$. That is, no residual interference will remain after retransmission, under some conditions of the fading gain interval. In addition to HARQ based protocols, we consider a retransmission scheme, which only utilizes the transmitter CSI on second transmission. The various protocols are numerically compared in section V.

The broadcast approach for the SISO channel was introduced in detail in [4]. For completeness of presentation we quickly review the principles of the broadcast approach. The incremental rate as function of power allocation is [4]

$$dR(u) = \frac{\rho(u)udu}{1 + I(u)u} \tag{21}$$

where $I(u)$ is the residual interference function, such that $I(0) = P$, and $\rho(u) = -\frac{d}{du}I(u)$ is the power density function. The maximal average rate ($M = 1$) is expressed as follows

$$R_{bs,avg} = \max E[R(s)] = \max_{I(u)} \int_0^\infty du(1 - F(u)) \frac{\rho(u)u}{1 + I(u)u} \tag{22}$$

where $F(u)$ is the fading gain cdf. The optimal power distribution, which maximizes (22), is [4]

$$I_{opt}(u) = \begin{cases} P & u < u_0 \\ \frac{1-F(u)-u\cdot f(u)}{u^2 f(u)} & u_0 \le u \le u_1 \\ 0 & u > u_1 \end{cases} \tag{23}$$

where $f(u)$ is the fading gain probability density function (pdf) of u, and the boundaries u_0 and u_1 are obtained from the boundary conditions $I_{opt}(u_0) = P$, and $I_{opt}(u_1) = 0$, respectively. This will also be the optimal power distribution for $M = 1$, however for other layered HARQ schemes it will probably be sub-optimal. Four Broadcast CC-HARQ (BCC-HARQ) and Broadcast IR-HARQ (BIR-HARQ) protocols are suggested.

A. BCC-HARQ, Protocol I

On the first transmission, the transmitter performs continuous broadcasting over $s \in [s_0, s_1]$, corresponding to the non-zero range of $\rho(s)$. Denote by $s_{eq}(s)$ the equivalent fading gain up to which decoding is possible on retransmission. Then, a retransmission ($M = 2$) is as follows:

1) $s < s_0$ - the receiver on the first transmission was in complete outage, retransmission is identical to first transmission (since no layer was reliably decoded).
2) $s \in [s_0, s_1)$, and $s_{eq}(s) < s_1$ - the receiver on the first transmission decoded layers up to $s > s_0$, however could not decode all layers. Retransmission ($M = 2$) consists of the layered information for undecoded layers, but **only** those layers which can be decoded at $M = 2$. That is, since the transmitter knows s, it can predict up to which layer the receiver will be able to decode. Therefore, retransmission with suitable power scaling is performed for layers in range $(s, s_{eq}(s)]$. The condition $s_{eq}(s) \le s_1$ suggests that not all originally transmitted layers can be decoded, and the equivalent fading gain representing the highest decodable layer $s_{eq}(s)$ is smaller than s_1.
3) $s \in [s_0, s_1)$, and $s_{eq}(s) = s_1$ - the receiver on the first transmission decoded layers up to $s_0 < s < s_1$, and can fully decode all layers after retransmission, since $s_{eq}(s) = s_1$. Then, retransmission consists of layered information for undecoded layers, and a new layer with new information. That is, since the transmitter knows s, which was sufficiently large, it designs the retransmission such that all originally transmitted layers will be decoded, and adapts the new layer such that zero outage is guaranteed.
4) $s \ge s_1$ - the receiver on the first transmission decoded all layers (no layer in outage). The retransmission ($M = 2$) includes only new information, in a rate corresponding to a fading gain s_1. On a first look, this may seem to be a sub-optimal approach, since the transmitter doesn't know s; it only knows $s \ge s_1$. However, it was shown in [21], that when $s_1 \ge s_1^{SISO}$, single level coding matched to s_1 is optimal, where s_1^{SISO} corresponds to s_1 dictated by an optimal broadcasting power allocation for a SISO channel with no transmit side information.

We now derive the rates, which express the throughput in Protocol-I. Consider a broadcasting power distribution $\rho(s)$, where $\rho(s) > 0$ for $s \in (s_0, s_1)$. The broadcasting residual interference function is defined as $I(s) = \int\limits_s^{s_1} \rho(u)du$. The average reward $R_{bs,(s<s_0)}$ corresponds to average achievable reward given $s < s_0$,

$$R_{bs,(s<s_0)} = \int\limits_{s_0/2}^{s_0} dsf(s) \int\limits_{s_0}^{\max\{s_0,2s\}} \frac{u\rho(u)du}{1 + uI(u)} \tag{24}$$

where $f(s)$ is the fading gain pdf, and $R(s) = \int\limits_{s_0}^{\max\{s_0,2s\}} du\frac{u\rho(u)du}{1+uI(u)}$ is obtained by optimally combining the two transmissions. The next achievable rate $R_{bs,(s_0\leq s<s_{eq}(s)<s_1)}$ is obtained for the case that not all layers are decoded even after the retransmission. The transmitter designs the retransmission in the following way. Power is allocated to undecoded layers and up to $s_{eq}(s)$. That is, all the transmitter available power is allocated to a subset of layers, which were not decoded on first transmission, and will be successfully decoded on retransmission:

$$\rho_2(u) = \begin{cases} \frac{P}{I(s)-I(s_{eq}(s))}\rho(u) & s \leq u \leq s_{eq}(s) \\ 0 & \text{otherwise} \end{cases}, \tag{25}$$

where $\rho(u)$ is the broadcasting power density during the first transmission. As may be noticed, the power allocation per layer only scales on retransmission, and the layers which cannot be decoded are not transmitted. Hence, after optimally combining the first and second transmissions, the fractional rate associated with fading gain s, is now

$$dR(s)|_{(s_0\leq s<s_{eq}(s)<s_1)} = \frac{(1+\mu(s))s\rho(s)ds}{1+(1+\mu(s))sI(s) - s\mu(s)I\left(s_{eq}(s)\right)} \tag{26}$$

where $\mu(s) = \frac{P}{I(s)-I(s_{eq}(s))}$. See Appendix A for a detailed derivation of (26). The fractional rate in (26) may be written in a form viewing an equivalent channel after the retransmission

$$dR_{bs,tot}(s)|_{(s_0\leq s<s_{eq}(s)<s_1)} = \frac{s_{eq}(s)\rho(s)ds}{1+s_{eq}(s)I(s)} \tag{27}$$

where some algebra on (26) leads to the following equivalent fading gain

$$s_{eq}(s)|_{s_{eq}(s)<s_1} = \frac{s(1+\mu(s))}{1-\mu(s)sI(s_{eq}(s))}. \tag{28}$$

For every s, the equivalent fading gain $s_{eq}(s)$ may be solved iteratively by substituting $\mu(s)$ into (28). The achievable broadcasting average reward is given by

$$R_{bs,(s_0\leq s<s_{eq}(s)<s_1)} = \int\limits_{s_0}^{s_b=s_{eq}^{-1}(s_1)} dsf(s) \int\limits_{s_0}^{s_{eq}(s)} du\frac{u\rho(u)du}{1+uI(u)} \tag{29}$$

where s_b is derived from (28). Note that $I(s_{eq}(s_b)) = 0$, which simplifies $\mu(s)$ and (28), hence

$$s_{eq}(s_b) = s_1 = s_b(1+\frac{P}{I(s_b)}) \tag{30}$$

from which s_b is extracted by solving the second equation. Consider now the case where the fading gain was $s_b < s < s_1$, which allows retransmission of all residual layers, and additional new information. In this case, retransmission power is divided between the residual layers (minimal value required for its reliable decoding), and a new layer, which rate is determined such that zero outage is guaranteed. The minimal power required for reliably decoding all layers, assuming $s > s_b$ is a modified version of (30), i.e. $s_1 = s(1 + \alpha(s)P/I(s))$, and then

$$\alpha(s) = \left(\frac{s_1}{s} - 1\right)\frac{I(s)}{P}. \tag{31}$$

For every $s_b < s < s_1$, $\alpha(s)P$ is allocated to residual layered data, and $(1-\alpha(s))P$ is allocated to a new layer to be decoded before the optimal combining of both transmissions. That is, the new information layer is decoded by considering the layered transmission as interference. After it is decoded, and removed from the retransmission signal, both transmissions are optimally combined to extract all decoded layers. The average reward here has a rather simple form since all layered data is completely decoded, and zero outage is achieved here,

$$R_{bs,(s_b<s<s_1,s_{eq}(s)=s_1)} = \int\limits_{s_b}^{s_1} dsf(s)\left\{\int\limits_{s_0}^{s_1}\frac{u\rho(u)du}{1+uI(u)} \ + \ \log\left(1+\frac{s(1-\alpha(s))P}{1+s\alpha(s)P}\right)\right\}. \tag{32}$$

The last case is when all layers were decoded on first transmission ($s \geq s_1$). Here, the retransmission includes only new data, which is transmitted as a single layer. That is,

$$R_{1L,(s\geq s_1)} = (1-F(s_1))\left(\int_{s_0}^{s_1}\frac{u\rho(u)du}{1+uI(u)} + \log(1+s_1P)\right). \tag{33}$$

The next proposition states the achievable rate result for this protocol.

Proposition 4.1: The average achievable throughput of BCC-HARQ with Protocol-I over a two block quasi-static fading channel (LTSC) is given by

$$\eta_{BCC,\text{Protocol-I}} = \frac{1}{2}\left(R_{bs,(s<s_0)} + R_{bs,(s_0\leq s<s_{eq}<s_1)} + R_{bs,(s_b<s<s_1,s_{eq}=s_1)} + R_{1L,(s\geq s_1)}\right) \tag{34}$$

where $E[D] = 2$, since $M = 2$, and a retransmission always occurs. The average rewards are specified by (24), (29), (32), and (33), respectively.

B. BCC-HARQ, Protocol II

On the first transmission, the transmitter performs continuous broadcasting over an interval $s \in [s_0, s_1]$, corresponding to the non-zero range of $\rho(s)$. Then retransmission ($M = 2$) follows:

1) $s < s_0$ - the receiver on the first transmission was in complete outage. **No retransmission** is performed. Thus an outage occurs on first transmission, however low latency of $D = 1$

is obtained, instead of low rate $D = 2$ in case of Protocol I. The motivation here is to increase efficiency by not using poor channel opportunities.

2) – 4) Same as in BCC-HARQ, Protocol I.

The achievable average rate is formulated in the following proposition.

Proposition 4.2: The average achievable throughput of BCC-HARQ with Protocol-II over a two block quasi-static fading channel (LTSC) is given by

$$\eta_{BCC,\text{Protocol-II}} = \frac{1}{2 - F(s_0)} \left(R_{bs,(s_0 \le s < s_{eq} < s_1)} + R_{bs,(s_0 \le s < s_1, s_{eq} = s_1)} + R_{1L,(s \ge s_1)} \right) \tag{35}$$

where $E[D] = 2 - F(s_0)$, and the average rewards are specified by (29), (32), and (33), respectively. The derivation of $E[D]$ is straightforward, by recalling that the only case with no retransmission is $s < s_0$, thus $E[D] = 1 + \Pr(s > s_0) = 2 - F(s_0)$.

C. Outage approach on retransmission (OAR)

On the first transmission, the transmitter performs continuous broadcasting over an interval $s \in [s_0, s_1]$, corresponding to the non-zero range of the power distribution $\rho(s)$, like in the previous protocols. However, on retransmission ($M = 2$) it sends only new information according to the following guidelines:

1) $s < s_0$ - the receiver on the first transmission was in complete outage. **No retransmission** is performed. Thus an outage occurs already on first transmission (like Protocol-II).
2) $s_0 \le s \le s_1$ - the second transmission includes only new data, in a rate matched to the index of the highest decoded layer (known from the feedback of first transmission). In this case there is no utilization of undecoded layered information received on first transmission.
3) $s \ge s_1$ - the receiver on the first transmission decoded all layers (no layer in outage). The retransmission ($M = 2$) includes only new information, in a rate corresponding to a fading gain s_1 (as done in the other protocols).

The achievable average rate is formulated in the following proposition.

Proposition 4.3: The average achievable throughput of the OAR protocol over a two block quasi-static fading channel (LTSC) is given by

$$\eta_{OAR} = \frac{1}{2 - F(s_0)} \left(R_{bs,SISO} + R_{erg,(s_0 \le s \le s_1)} + R_{1L,(s \ge s_1)} \right) \tag{36}$$

where $E[D] = 2 - F(s_0)$*, and the average rewards are specified by*

$$R_{bs,SISO} = \int_{s_0}^{s_1} dsf(s) \int_{s_0}^{s} \frac{u\rho(u)du}{1 + uI(u)} \tag{37}$$

$$R_{erg,(s_0 \leq s \leq s_1)} = \int_{s_0}^{s_1} dsf(s) \log(1 + sP), \tag{38}$$

and $R_{1L,(s \geq s_1)}$ *is given by (33).*

Interestingly, the optimal power distribution for this protocol is immediately given by the SISO optimal power distribution of $M = 1$.

Proposition 4.4: Optimal power allocation for OAR is given by $I_{opt,\text{OAR}}(s) = I_{opt}(s)$ *in (23).*

Proof: The subject for optimization in η_{OAR} is $I(s)$. This optimization can be explicitly written as the following variational problem

$$\eta_{OAR} = \max_{I(s), s \geq 0} \frac{1}{2 - F(s_0)} \left(\int_{s_0}^{s_1} ds \left[(1 - F(s)) \frac{u\rho(u)du}{1 + uI(u)} + f(s) \log(1 + sP) \right] + \text{const}(s_1) \right) \tag{39}$$

where we have used partial integration for $R_{bs,SISO}$, and const(s_1) is an independent constant. As may be observed from (39), only the first term depends on $I(s)$, and $I'(s) = -\rho(s)$. This term constructs $R_{bs,SISO}$. Hence the extremum condition yielding $I_{opt,\text{OAR}}(s)$ is identical to that corresponding to SISO broadcasting with $M = 1$. ■

D. Broadcast IR (BIR)-HARQ Protocol

We focus here on the BIR-HARQ with $M = 2$. On the first transmission, the transmitter performs continuous broadcasting over an interval $s \in [s_0, s_1]$, corresponding to the non-zero range of $\rho(s)$. On the retransmission, the strategy considered is similar to that of BCC-HARQ, Protocol-II. The BIR-HARQ is different from the BCC-HARQ protocol only in the retransmission for $s \in [s_0, s_1)$. In other cases the BIR-HARQ retransmission is identical to the BCC-HARQ retransmission scheme.

In the BIR-HARQ scheme the retransmission includes IR information in a layered manner, to undecoded layers only. Each layer receives its own IR data. The retransmission is jointly decoded with first transmission governed by the sum mutual information of the two transmissions. The next proposition defines the achievable rate of a BIR-HARQ approach.

Proposition 4.5: The rate achievable with BIR-HARQ with $M = 2$ is given by

$$\frac{s_{eq}\rho(s_{eq})ds_{eq}}{1 + s_{eq}I(s_{eq})} = \frac{s\rho(s_{eq})ds_{eq}}{1 + sI(s_{eq})} + s\mu\rho(s_{eq})ds_{eq} \tag{40}$$

with

$$s_{eq} = \frac{s(1+\mu) + s^2\mu I(s_{eq})}{1 - s\mu(1 + sI(s_{eq}))I(s_{eq})} \tag{41}$$

where for notational brevity we have replaced $s_{eq}(s)$ by s_{eq}, and $\mu(s)$ by μ. The solution of $s_{eq}(s)$ in (41) is iteratively computed for a given $I(u)$, with $\mu(s) = \frac{\alpha(s)P}{I(s)-I(s_{eq}(s))}$.

Proof: We start by showing that for every $u \in [s, s_{eq}(s)]$ the following inequality is satisfied

$$\frac{u\rho(u)du}{1 + uI(u)} \leq \frac{s\rho(u)du}{1 + sI(u)} + \frac{s\mu\rho(u)du}{1 + s\mu(I(u) - I(s_{eq}))} \tag{42}$$

where the expressions on the right hand side correspond to the mutual information associated with layer u from first and second transmissions, respectively. After some algebra, the above simplifies into

$$\frac{u}{1 + uI(u)} \leq \frac{s}{1 + sI(u)} + \frac{s_x}{1 + s_xI(u)} \tag{43}$$

where $s_x \triangleq \frac{s\mu}{1-s\mu I(s_{eq})}$. For $u = s$ there is a strict inequality, and the functions on both sides of the inequality are monotonically increasing functions of u. This is since $\frac{dI(u)}{du} = -\rho(u)$, and $\rho(u) > 0$ in the range $u \in [s, s_{eq}(s)]$. Define $G(u) = \frac{u}{1+uI(u)}$, then $\frac{dG(u)}{du} = \frac{1+u^2\rho(u)}{[1+uI(u)]^2}$, and for the same reason, $\frac{dG(u)}{du} > 0$ for $u \in [s, s_{eq}(s)]$. As $G(u)$ increases faster than the RHS of (43), we denote by $u = s_{eq}$ the point where the inequality becomes an equality. When substituting $u = s_{eq}$ into $\frac{s_x}{1+s_xI(u)}$, we get $\frac{\frac{s\mu}{1-s\mu I(s_{eq})}}{1+\frac{s\mu}{1-s\mu I(s_{eq})}I(s_{eq})} = s\mu$. And by requiring equality in (43), the highest fractional rate of the highest decodable layer is given by (40). And $s_{eq}(s)$ (41) is an immediate algebraic derivation of (40). ■

In order to determine s_b, like in BCC-HARQ, it is required to derive $s_b = s_{eq}^{-1}(s_1)$. This is directly derived from (41), by substituting s_{eq} by s_1. This results in

$$s_b = \frac{s_1}{1 + \mu(s_b)} = \frac{s_1}{1 + \frac{P}{I(s_b)}}, \tag{44}$$

which is the same case as in BCC-HARQ (30). Naturally, the power allocation $\alpha(s)$ will be the same as specified in (31) for $s_b < s \leq s_1$, since it is derived from (44).

Since s_b is identical for BCC-HARQ and BIR-HARQ, a question arises from both approaches. Is BIR better than BCC? The next proposition shows that for the interesting case of $s \in [s_0, s_1]$, the BIR outperforms BCC.

Proposition 4.6: $s_{eq,BIR}(s) > s_{eq,BCC}(s)$ *- The equivalent fading gain, which corresponds to highest decodable layer, for a BIR-HARQ approach is greater than that associated with a BCC-HARQ approach, for* $s \in (s_0, s_b)$*, and a given* $I(s)$.

Proof: Let us first explicitly state $s_{eq,BIR}(s)$ (41), and $s_{eq,BCC}(s)$ (28),

$$s_{eq,BIR} = \frac{s(1+\mu) + s^2\mu I(s_{eq,BIR})}{1 - s\mu(1 + sI(s_{eq,BIR}))I(s_{eq,BIR})} \tag{45}$$

and

$$s_{eq,BCC} = \frac{s(1+\mu)}{1 - s\mu I(s_{eq,BCC})}. \tag{46}$$

Using a perturbation approach, assume $s_{eq,BIR} = s_{eq,BCC} = u$. The following has to be shown

$$\frac{s(1+\mu)}{1 - s\mu I(u)} \lesseqgtr \frac{s(1+\mu) + s^2\mu I(u)}{1 - s\mu(1 + sI(u))I(u)}, \tag{47}$$

for u. By simplifying the right hand side of (47) the inequality is evident,

$$\frac{s(1+\mu)}{1 - s\mu I(u)} < \frac{s(1+\mu) + s^2\mu I(u)}{1 - s\mu I(u) - s^2\mu I(u)^2} \tag{48}$$

where the strict inequality holds for $0 < s < u$, and $I(u) > 0$. Since both functions are monotonically increasing w.r.t u, it is clear that $s_{eq,BIR}(s) > s_{eq,BCC}(s)$. ■

V. Numerical Results

We present here numerical results for the IR with finite level multi-layer coding, and the various continuous broadcasting protocols. The results include achievable rates for classical IR-HARQ, with single level coding. These are compared to two level coding IR-HARQ as derived in the section III. Continuous broadcast HARQ refers to both BCC-HARQ and BIR-HARQ for $M = 2$, as derived in section IV.

Figure 1, demonstrates the achievable throughput of the outage and two level coding IR-HARQ approaches under the LTSC model. The results are compared to the outage lower bound (M=1), and the ergodic capacity upper bound. Notice that the highest broadcasting gain is achieved for $M = 2$. That is, if only one retransmission is allowed, then the broadcasting gain over the classical IR-HARQ (employing an outage approach) is largest. In other cases ($M \geq 4$) the performance of both approaches nearly match, from which we conjecture that as M grows the broadcasting gain vanishes, and outage approach becomes optimal. We note though that accurate numerical results for achievable throughput with two level coding may be obtained only for high SNRs, due to the high SNR approximation in $p_{o,2}(m)$ (15).

Figures 2-3, demonstrate the achievable throughput as function of R in the LTSC model. Note that for two level coding the x-axis represents the sum-rate $R = R_1 + R_2$. And the achievable rate presented is for the optimal rate and power allocation for each sum-rate. This also explains why in both figures, as R grows, the HARQ rates converge to a constant rate. This is since the same fixed rate MR_1 is always chosen, with $\alpha = 1$, and a constant decoding delay of M. The residual rate for R_2 is never decoded (since it receives zero power), hence the latency penalty is always M.

Figure 4 demonstrates the achievable rates for two level coding, with only one retransmission ($M = 2$). The two channel models are considered here, namely the LTSC and STSC models. As expected, the layering over an STSC is more efficient than over an LTSC. It is expected that as M increases, the layering-IR over STSC will approach the ergodic capacity faster than over the LTSC.

Figures 5-6, demonstrate the achievable throughput of the BIR-HARQ strategy corresponding to Protocols I-II, and the outage approach on retransmission (OAR) scheme, which were all defined in Section IV. Since the optimal power distribution for continuous broadcasting was not obtained, we turn to sub-optimal power distributions, and focus on two broadcasting power distributions: $I_{SISO,opt}(s)$ - refers to the optimal broadcasting power distribution specified in (23) for a SISO channel with $M = 1$, which is also an optimal distribution for the OAR approach; $I_{SIMO,opt}(s)$ - refers to the optimal broadcasting power distribution specified in (23) for a SIMO channel with two receive antennas. As may be noticed from Figure 5, in the low-moderate SNR range, BIR-HARQ-II outperforms the other protocols, with $I_{SIMO,opt}(s)$ power distribution. Notice there a $\sim$3 dB gain of BIR-HARQ over the classical broadcast approach with $M = 1$. It even outperforms the finite level coding IR with $M = 10$, as presented in Figure 1. Figure 6, shows that for high SNRs OAR outperforms the others, which means that for high SNRs it is better to sends new information rather than a layered chase combining or IR retransmission. Table I demonstrates the small (even negligible) gain in BIR-HARQ over BCC-HARQ. However, there is no immediate conclusion to draw here about BIR/BCC-HARQ, as the broadcasting power distribution used is sub-optimal.

We note here that the BCC/BIR-HARQ achievable rates were obtained using sub-optimal broadcasting power distributions. From the corresponding numerical results it seems very appealing to use the OAR scheme for its low implementation complexity, and since it is the most efficient scheme for high SNRs, and it closely approximates the BIR-HARQ for low SNRs. However, finding the broadcasting optimal power distribution of BIR-HARQ is still an open problem. It may as well turn out that BIR-HARQ with an optimal power distribution has a more pronounced gain over the OAR scheme, for which the optimal power allocation was fully characterized here. Generally, the BIR-HARQ is expected to outperform the OAR scheme since having the broadcast approach used on the first transmission, facilitates not only to have an efficient scheme which competes with having the CSI beforehand, and that is because once the CSI is accurately known via the feedback (in the continuous layering case), then for all the layers that should be decoded in the second round only the additional information (not wasting that accumulated in the first round) is sent.

From Figure 7, it may be noticed that the broadcasting high throughput with delay of two blocks is nearly achieved for conventional HARQ of $M = 10$, with an average delay greater than 6 blocks. Figure 8 shows that high BIR-HARQ gains are obtained for a wide SNR range.

Figure 9 focuses on the comparison of BIR-HARQ and two level coding. When $M = 1$ is the delay constraint, i.e. no HARQ is used, it is known that two level coding closely approximates the continuous broadcasting upper bound [4]. However, with the BIR-HARQ this is not the case anymore, like demonstrated in Figure 9. The gain of BIR-HARQ over two level coding is ∼4 dB for equal delay ($M = 2$). Even for $M = 4$, two level coding is ∼2 dB far from the low-delay BIR-HARQ ($M = 2$) scheme.

VI. Conclusion

We have studied various multi-layer encoding HARQ schemes. The motivation for extending the conventional HARQ schemes to multi-layer coding is to achieve high throughput efficiencies with low latency. The study focused on finite level coding IR-HARQ, where every code layer supports IR coding. The multi-layer bounds were investigated through continuous broadcasting, by defining different protocols for BCC-HARQ and BIR-HARQ. An optimal power distribution cannot be obtained for continuous broadcasting. However, it was observed that even with a sub-optimal broadcasting power distribution pronounced gains of ~ 3 dB over an outage approach, can be achieved for low and moderate SNRs, in the LTSC model, with a very low latency of two blocks. This is especially interesting, as the conventional broadcast approach (without HARQ), has only marginal gains over the outage approach for low SNRs.

The OAR protocol is also an interesting approach, which uses retransmissions for sending new information, in a rate matched to the broadcasting feedback from first transmission. The optimal broadcasting power distribution for OAR was fully characterized here, and numerical results showed it is the most efficient scheme for high SNRs, and it closely approximates the BIR-HARQ for low SNRs. However, in BIR-HARQ only sub-optimal power distributions were used, and finding the broadcasting optimal power distribution is still an open problem. It may as well turn out that the BIR-HARQ with an optimal power distribution has more pronounced gains over the OAR scheme.

APPENDIX A
DETAILED DERIVATION OF $dR(s)$ IN (26)

In order to clearly specify the fractional rate associated with the fading gain s following the second transmission, the following representation of the received signals may be helpful:

$$\mathbf{y}_1 = h\mathbf{x}_s + h\mathbf{x}_I + \mathbf{n}_1 \tag{A.1}$$

where $\mathbf{y}_1$ is the received vector in first transmission, and the originally transmitted signal is $\mathbf{x} = \mathbf{x}_s + \mathbf{x}_I$, where $\mathbf{x}_s$ represents the decodable layered data, and the residual interference is denoted by $\mathbf{x}_I$, with power $I(s)$. The second transmission is carefully designed such that only the jointly decodable part of $\mathbf{x}_I$ is transmitted, thus

$$\mathbf{y}_2 = h\sqrt{\mu(s)} \cdot \mathbf{x}_{Is} + \mathbf{n}_2 \tag{A.2}$$

where $\sqrt{\mu(s)}$ is a power normalization factor, specified below (26), and $\mathbf{x}_{Is}$ denotes the jointly decodable part of $\mathbf{x}_I$. Thus $\mathbf{x}_I$ can be written as $\mathbf{x}_I = \mathbf{x}_{Is}+\mathbf{x}_{In}$, where $\mathbf{x}_{In}$ is a residual interference following the second transmission, and its power is $I(s_{eq}(s))$ where $s_{eq}(s)$ is associated with the highest decodable layer after optimally processing the two transmissions. Hence the received signal in (A.1), after removal of $\mathbf{x}_s$, can be written as $\mathbf{y}_{1,c} = h\mathbf{x}_{Is}+h\mathbf{x}_{In}+\mathbf{n}_2$, and after optimally combining $\mathbf{y}_2$ and $\mathbf{y}_{1,c}$ we get

$$\mathbf{y}_c = \mathbf{x}_{Is} \cdot s(1+\mu(s)) + s\mathbf{x}_{In} + h\mathbf{n}_1 + h\sqrt{\mu(s)} \cdot \mathbf{n}_2. \tag{A.3}$$

From (A.3), the fractional rate for decoding sub-layer s is given by

$$dR(s) = \frac{(1+\mu(s))s\rho(s)ds}{1+sI(s)+s\mu(s)(I(s)-I(s_{eq}(s)))} \tag{A.4}$$

where the first $I(s)$ in the denominator amounts to residual interference in $\mathbf{y}_{1,c}$, and $I(s) - I(s_{eq}(s))$ is the residual interference contributed by $\mathbf{y}_2$. Subtraction of $I(s_{eq}(s))$ is the result of the careful power allocation for retransmission (25). From (A.4), derivation of (26) is straightforward.

References

[1] L. Ozarow, S. Shamai (Shitz), and A. Wyner, "Information theoretic considerations for cellular mobile radio," *IEEE Trans. Veh. Tech.*, vol. 43, no. 2, pp. 359–378, May 1994.

[2] D. Chase, "Code combining–a maximum-likelihood decoding approach for combining an arbitrary number of noisy packets," *IEEE Transactions on Communications*, vol. 33, no. 5, pp. 385–393, May 1985.

[3] S. Shamai (Shitz), "A broadcast strategy for the Gaussian slowly fading channel," *IEEE ISIT'97, Ulm Germany*, p. 150, June 29–July 4 1997.

[4] S. Shamai (Shitz) and A. Steiner, "A broadcast approach for a single user slowly fading MIMO channel," *IEEE Trans. on Info, Theory*, vol. 49, no. 10, pp. 2617–2635, Oct. 2003.

[5] T. Cover, "Broadcast channels," *IEEE Trans. on Info. Theory*, vol. 18, no. 1, pp. 2–14, Jan. 1972.

[6] G. Caire and D. Tuninetti, "The throughput of hybrid-ARQ protocols for the Gaussian collision channel," *IEEE Transactions on Information Theory*, vol. 47, no. 5, pp. 1971–1988, July 2001.

[7] S. Sesia, G. Caire, and G. Vivier, "Incremental redundancy hybrid ARQ schemes based on low-density parity-check codes," *IEEE Transactions on Communications*, vol. 52, no. 8, pp. 1311–1321, August 2004.

[8] N. Varnica, E. Soljanin, and P. Whiting, "LDPC code ensembles for incremental redundancy hybrid ARQ," *in IEEE Int. Symp. Inform. Theory (ISIT'05), Adelaide, Australia*, pp. 995–999, Sept. 4-9 2005.

[9] J. Kim, W. Hur, A. Ramamoorthy, and S. McLaughlin, "Design of rate-compatible irregular LDPC codes for incremental redundancy hybrid ARQ systems," *in IEEE Int. Symp. Inform. Theory (ISIT'06), Seattle, WA*, July 9-14 2006.

[10] E. Soijanin, N. Varnica, and P. Whiting, "Punctured vs rateless codes for hybrid ARQ," *in IEEE Information Theory Workshop (ITW'06), Punta del Este, Uruguay*, pp. 155–159, March 13-17 2006.

[11] W. Hur and S. W. McLaughlin, "Incremental redundancy low-density parity check codes for MIMO V-BLAST systems," *in Proc. 40th Conference on Information Sciences and Systems (CISS'06), Princeton NJ*, March 2006.

[12] R. Liu, P. Spasojevic, and E. Soljanin, "A throughput analysis of incremental redundancy hybrid ARQ schemes with turbo codes," *in Proc. Conference on Information Sciences and Systems (CISS'04), Princeton NJ*, March 2004.

[13] ——, "Cooperative diversity with incremental redundancy turbo coding for quasi-static wireless networks," *in Proc. the 6th IEEE International Workshop on Signal Processing Advances for Wireless Communications, NYC*, June 2005.

[14] N. Dutsch and J. Hagenauer, "Combined incremental and decremental redundancy in joint source-channel coding," *in Proc. International Symposium on Information Theory and its Applications (ISITA'04), Parma, Italy*, pp. 775–779, October 2004.

[15] Z. Yiqing and W. Jiangzhou, "Optimum subpacket transmission for hybrid ARQ systems," *IEEE Trans. on Communications*, vol. 54, no. 5, pp. 934–942, May 2006.

[16] J. Roberson and D. Zhi, "A BICM approach to type-II hybrid ARQ," *in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'06), Toulouse, France*, vol. 4, pp. 273–276, May 15-19 2006.

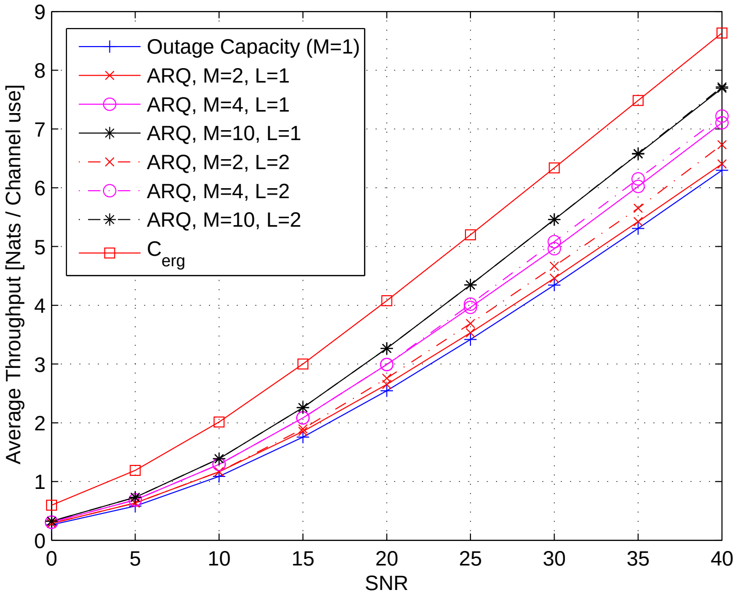

Fig. 1. Achievable throughput with different IR-HARQ schemes for the LTSC model. The classical outage approach serves as the lower bound, and the ergodic capacity serves as upper bound. Solid lines refer to classical IR-HARQ schemes (L=1), and in dashed lines the two level coding with IR-HARQ (L=2).

[17] H. El Gamal, G. Caire, and M. O. Damen, “The MIMO ARQ channel: Diversity-multiplexing-delay tradeoff,” *IEEE Transactions on Information Theory*, vol. 52, no. 8, pp. 3601–3621, August 2006.

[18] T. Tabet, S. Dusad, and R. Knopp, “Achievable diversity–multiplexing–delay tradeoff in half-duplex ARQ relay channels,” *in IEEE Int. Symp. Inform. Theory (ISIT’05), Adelaide, Australia*, pp. 1828– 1832, Sept. 4-9 2005.

[19] I. Stanojev, O. Simeone, and Y. Bar-Ness, “Performance analysis of collaborative hybrid-ARQ protocols over fading channels,” *in Proc. IEEE Sarnoff Symposium, Princeton, NJ*, March 2006.

[20] M. Zorzi and R. R. Rao, “On the use of renewal theory in the analysis of ARQ protocols,” *IEEE Transactions on Comm.*, vol. 44, no. 9, pp. 1077–1081, September 1996.

[21] A. Steiner and S. Shamai (Shitz), “Broadcasting with patial transmit channel state information,” in *NEWCOM–ACoRN Joint workshop 2006*, Vienna, Austria, September 20-22 2006.

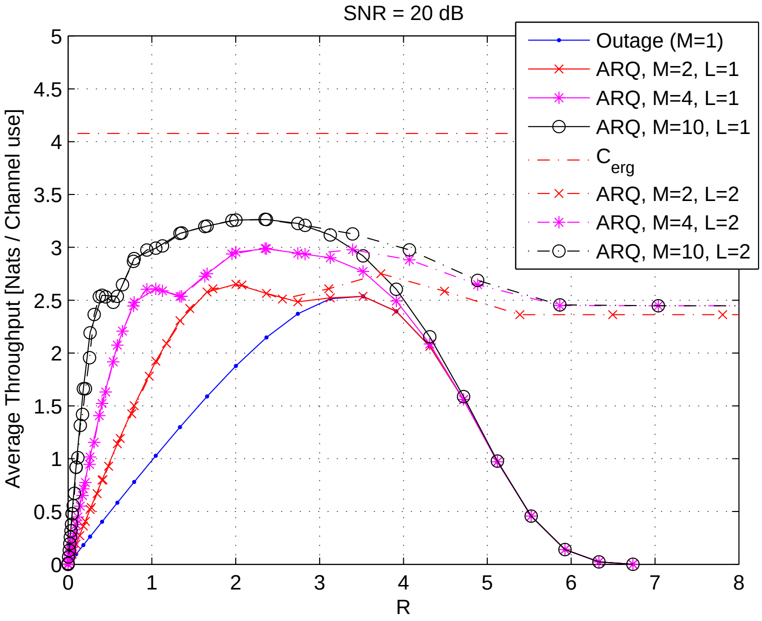

Fig. 2. Achievable throughput with different IR-HARQ schemes for the LTSC model. The classical outage approach serves as the lower bound, and the ergodic capacity serves as upper bound. Solid lines refer to classical IR-HARQ schemes (L=1), and in dashed lines the two level coding with IR-HARQ (L=2). SNR=20dB.

SNR	0	10	20	30	40
BIR-HARQ	0.444	1.628	3.398	5.488	7.704
BCC-HARQ (II)	0.444	1.627	3.395	5.486	7.703

TABLE I

ACHIEVABLE BIR-HARQ RATES VERSUS BCC-HARQ (PROTOCOL II) RATES, WITH $M = 2$. THE TABLE INCLUDES HIGHEST ACHIEVABLE RATE WITH EITHER $I_{SISO,opt}$ OR $I_{SIMO,opt}$.

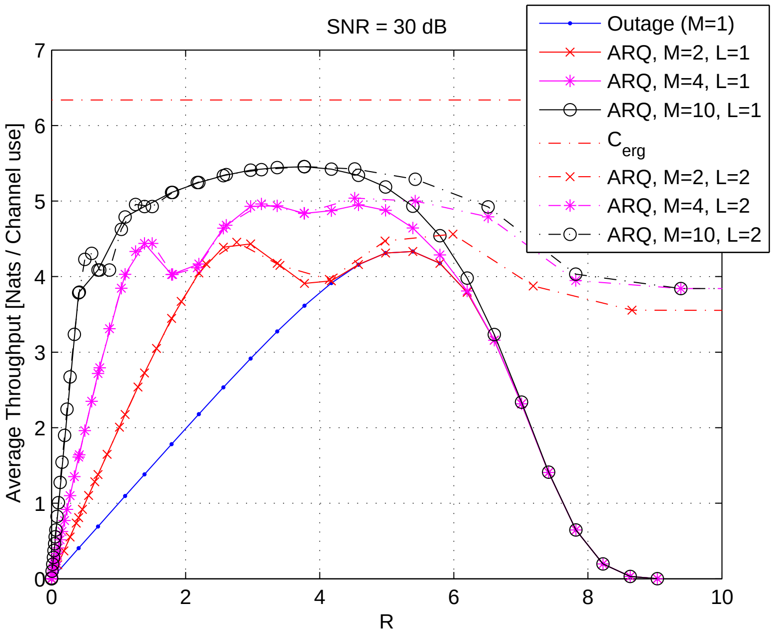

Fig. 3. Achievable throughput with different IR-HARQ schemes for the LTSC model. The classical outage approach serves as the lower bound, and the ergodic capacity serves as upper bound. Solid lines refer to classical IR-HARQ schemes (L=1), and in dashed lines the two level coding with IR-HARQ (L=2). SNR=30dB.

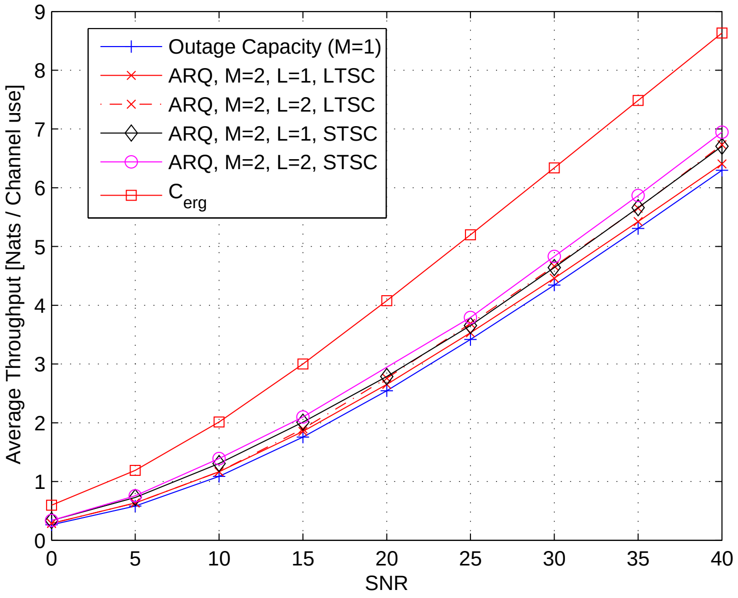

Fig. 4. Achievable throughput with different IR-HARQ schemes. The classical outage approach serves as the lower bound, and the ergodic capacity serves as upper bound. The single level coding is compared to two-level coding ($L = 2$), under two channel models (the LTSC and STSC models).

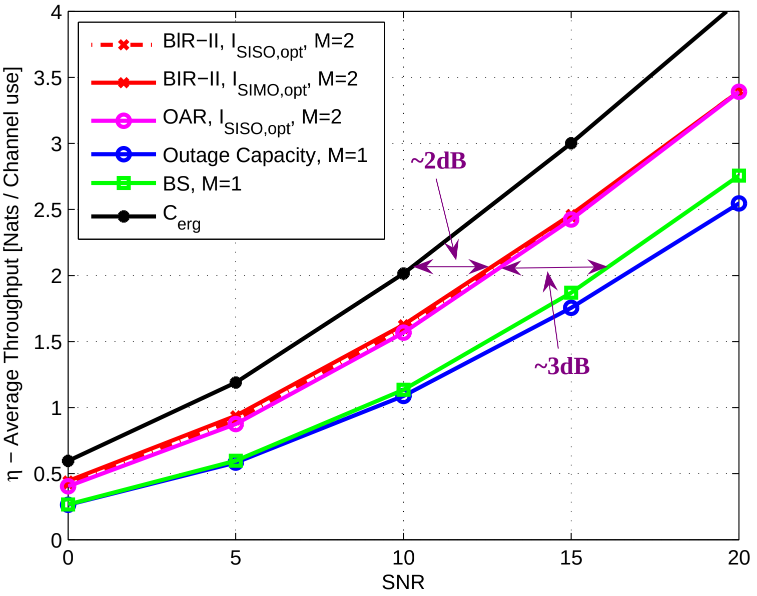

Fig. 5. Low SNR achievable throughput with BIR and OAR protocols ($M = 2$). The classical outage approach serves as the lower bound, and the ergodic capacity serves as upper bound. $I_{SISO,opt}$ refers to the optimal broadcasting power distribution of a SISO channel with $M = 1$. $I_{SIMO,opt}$ refers to the optimal broadcasting power distribution of a SIMO channel with two receive antennas, and with $M = 1$.

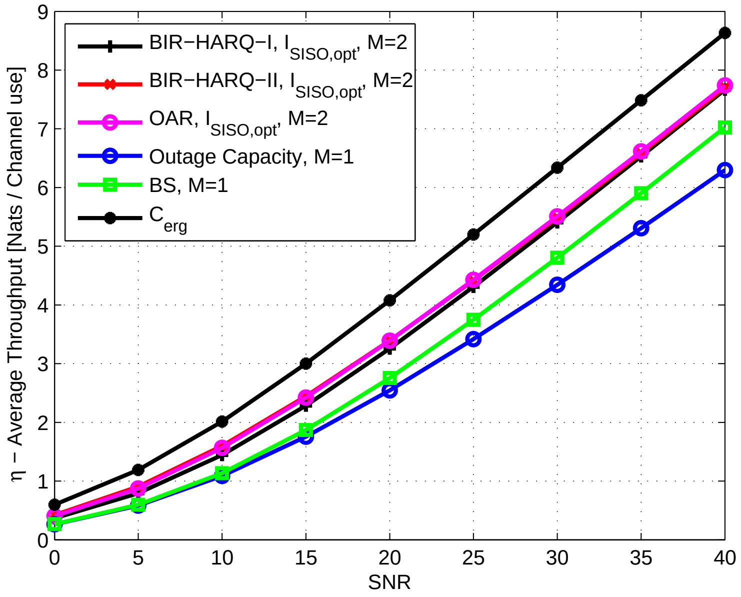

Fig. 6. Achievable throughput with BIR and OAR protocols ($M = 2$). The classical outage approach serves as the lower bound, and the ergodic capacity serves as upper bound. $I_{SISO,opt}$ refers to the optimal broadcasting power distribution of a SISO channel with $M = 1$.

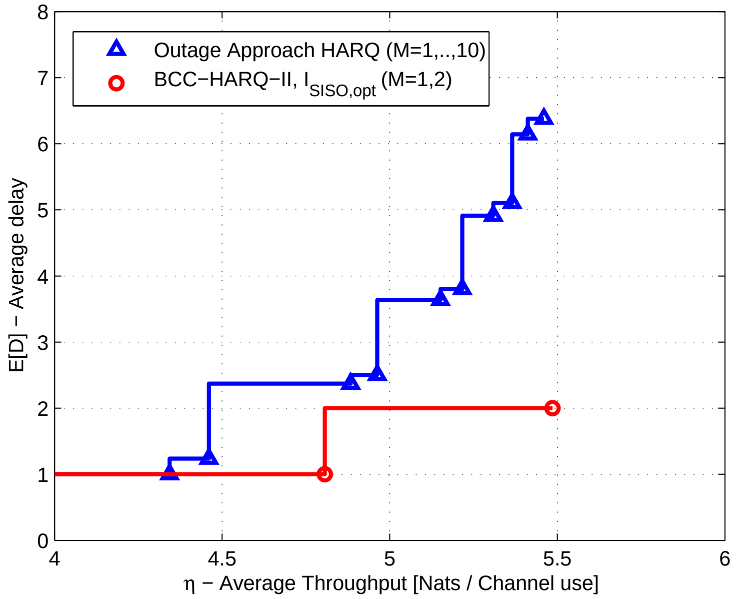

Fig. 7. Average delay as function of the average achievable throughput. The outage approach HARQ schemes with $M = 1, 2, ..., 10$ are compared to the broadcasting HARQ, where for $E[D] = 1$ the SISO broadcasting is used, and for $E[D] = 2$ the BIR protocols II is used ($P = 30$ dB).

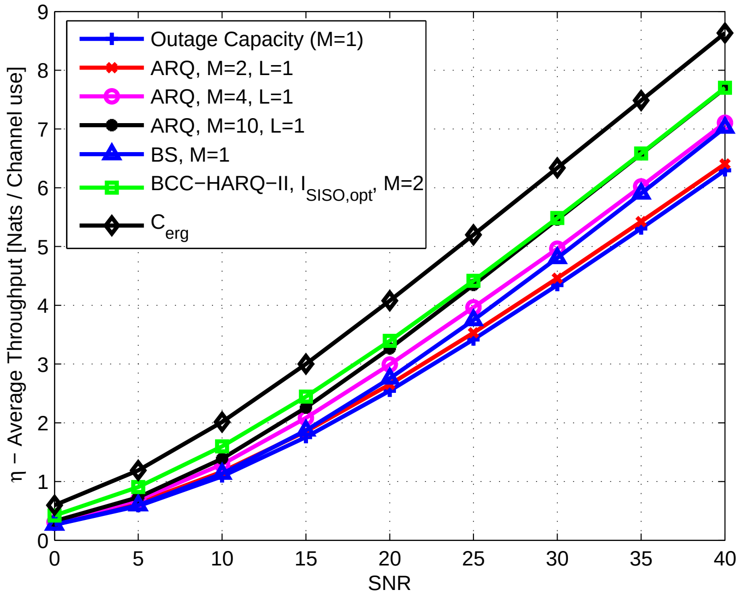

Fig. 8. Achievable throughput with BIR protocol II ($M = 2$), and the conventional outage HARQ schemes ($M = 2, 4, 10$).

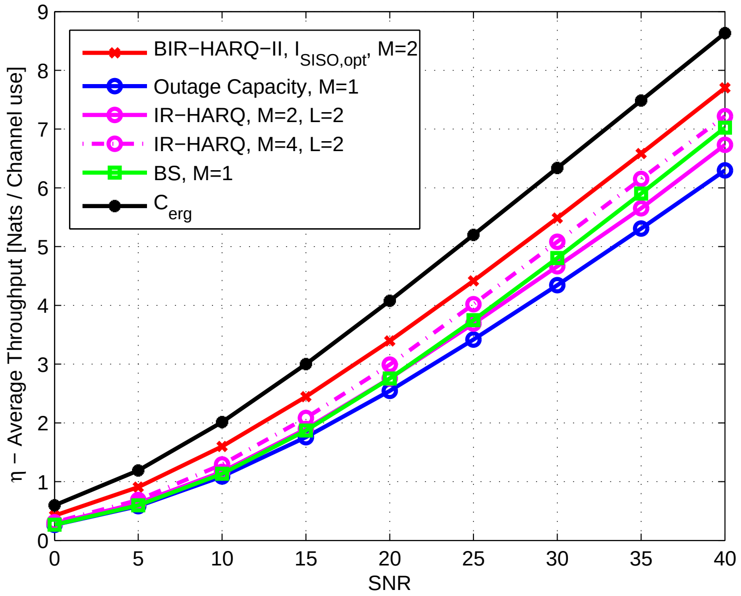

Fig. 9. Achievable throughput with BIR protocols ($M = 2$), compared to two-level coding with $M = 2, 4$. $I_{SISO,opt}$ refers to the optimal broadcasting power distribution of a SISO channel with $M = 1$.