Multi-Packet HARQ with Delayed Feedback

Alaa Khreis, Philippe Ciblat, Francesca Bassi, Pierre Duhamel

Multi-Packet HARQ with Delayed Feedback

Alaa Khreis[1], Philippe Ciblat[1], Francesca Bassi[2,3], and Pierre Duhamel[2]
[1] LTCI, Telecom ParisTech, Université Paris-Saclay, F-75013 Paris, France
[2] L2S, Université Paris-Saclay, F-91192 Gif-sur-Yvette, France
[3] ESME-Sudria, F-94200 Ivry-sur-Seine, France
Email: {alaa.khreis,ciblat}@telecom-paristech.fr, {bassi,duhamel}@l2s.centralesupelec.fr

***Abstract*—In current wireless communication systems, the feedback required by the Hybrid Automatic ReQuest (HARQ) mechanism is received with some delay at the transmitter side. To alleviate this issue, parallel Stop-and-Wait HARQ is usually employed. In this paper, we propose a multi-packet HARQ protocol (also called superposition coding or multi-layer HARQ) to improve the user's delay distribution and increase the throughput, without any additional feedback such as Channel State Information. The performance analysis, provided from an information-theoretic point-of-view, shows that the proposed protocol offers better delay distribution, higher throughput and lower message error rate compared to the conventional parallel Stop-and-Wait HARQ, at the expense of increased decoding complexity.**

I. Introduction

Hybrid Automatic Repeat ReQuest (HARQ) with Incremental Redundancy (IR) is used in wireless communication systems to improve communication reliability [1]. In HARQ, the receiver feeds back an Acknowledgment (ACK) (resp. an Negative ACK) to the transmitter if the receiver succeeds (resp. fails) to decode the current message. If ACK is fed back, the transmitter sends a packet related to a new message during the next time-slot. If NACK is fed back and the maximum retransmission credit is not exhausted, the transmitter sends a redundant packet associated to the current message.

Due to propagation time and reverse link scheduling, the feedback associated to a sent packet is available to the transmitter with a delay of multiple time-slots. The communication system is idle in-between if Stop-and-Wait HARQ protocol is used [2]. A conventional approach to compensate for this issue is to initiate, during the unused time-slots, Stop-and-Wait HARQ processes corresponding to other messages. This scheme, called parallel Stop-and-Wait HARQ [2], allows the transmission of multiple messages in parallel, each of them employing an independent Stop-and-Wait HARQ process. Parallel Stop-and-Wait HARQ improves the throughput but does not mitigate the delay of the received messages, which is related to the maximum retransmission credit and to the feedback delay. The protocol proposed in this paper enables the transmitter to anticipate the feedback by sending, in advance to its reception, data related to unacknowledged messages. This is accomplished using superposition of packets, also called multi-packet or multi-layer transmission in the literature [3]–[8].

To the best of our knowledge, none of the previous works on superposition coding in the context of HARQ considered the possibility of delayed feedback. Therefore the proposed protocol, described in Section III, is significantly novel since it allows to counteract the delayed feedback as well as to improve the throughput. In [3], [4], superposition coding without feedback delay is employed, and outdated Channel State Information (CSI) is available at the transmitter side to optimize the choice of superposed packets. The protocol in [5] considers feedback on the Transport Control Protocol (TCP) layer, where each TCP frame corresponds to a pre-defined number of retransmissions with superposed packets. In [6], the authors propose a simple superposition modulation based on QPSK (Quadrature Phase-shift Keying), with a single possible retransmission and without feedback delay. In [7], the authors focus on practical joint detection of superposed packets in multi-layer HARQ without feedback delay. In [8], the authors show that multi-layer HARQ (considered there with no feedback delay) significantly increases the throughput.

An alternative method to improve the conventional parallel HARQ is to use time sharing and rate adaption policies, as in [9] and [10]. This requires, however, knowledge of the CSI at the transmitter side. Moreover, time-sharing does not exploit the potential of Multiple Access Channel (MAC) communication, while superposition does.

The main contribution of this work is a protocol that superposes an additional layer of redundant packets to the parallel Stop-and-Wait HARQ protocol. In the superposed layer, the transmitter may perform retransmissions of a message even before having received any feedback about it. The selection of the superposed packets is based solely on the delayed ACK/NACK feedback (no additional CSI). The proposed protocol improves the performance of parallel Stop-and-Wait HARQ in terms of delay distribution, throughput and message error rate at the expense of decoding complexity, since the receiver must decode superposed packets.

This work was supported by the Labex Digicosme PhD scholarship from Université Paris-Saclay under the grant called "Coccinelle"

This paper is organized as follows: Section II defines the system model. We explain the proposed HARQ protocol in Section III and the corresponding receiver is analyzed from an information theoretic point of view in Section IV. Performance metrics in comparison to conventional HARQ are presented in Section V. Concluding remarks are drawn in Section VI.

II. System model

We consider slotted point-to-point transmission where each time-slot corresponds to N channel uses. During each time-slot t, the transmitter sends N symbols stacked in vector $\mathbf{x}_t$. This vector may be composed of a packet or a superposition of packets, as it will be explained in Section III. The received signal at time-slot t is

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

where $\mathbf{n}_t$ is an additive white Gaussian noise vector, with zero-mean and variance per component equal to N_0. We consider a Rayleigh flat fading channel with coherence time equal to the time-slot duration. The channel is fixed during a time-slot, but has independent realizations at each time-slot. We also assume perfect CSI at the receiver. The channel gain is denoted by g_t where $g_t = \frac{|h_t|^2}{N_0}$. The transmitter has a set of messages $\{\mathbf{m}_k\}_{k\in\mathbb{N}_+}$ to send. Each message contains NR bits of information, where R is the rate. A message $\mathbf{m}_k$ is encoded via a mother code of rate $R_0 \in \mathbb{R}_+$ and then punctured into C modulated packets of length N. Consequently, we have $R = R_0 C$. The ℓ-th packet related to message $\mathbf{m}_k$ is denoted by $\mathbf{p}_k(\ell)$ with $\ell \in \{1, \cdots, C\}$. The feedback is error-free and only composed of ACK or NACK of the considered messages. We assume a feedback delay of T time-slots, which means that the feedback related to a transmission performed in time-slot t is received by the transmitter just before the beginning of time-slot $t+T$. The case $T = 1$ corresponds then to a no-delay feedback. We assume moreover that this delay is due to the return channel and not to the decoding time at the receiver, which means that the receiver knows at the end of time-slot t if the messages related to the packets transmitted at time-slot t are successfully decoded or not. A message is said in *timeout* if it is not ACKed by CT time-slots after its first transmission, corresponding to the timeout in conventional parallel HARQ.

III. Proposed Protocol

In the proposed protocol, at each time-slot the transmitter selects a packet $\mathbf{p}_k(\ell)$, based on the ACK/NACK feedbacks, as in conventional parallel Stop-and-Wait HARQ. The transmitter may superpose to $\mathbf{p}_k(\ell)$ a second packet $\mathbf{p}_{k'}(\ell')$, with $k' \neq k$, even if there is not any feedback on previous transmissions of message $\mathbf{m}_{k'}$ yet. The idea is to send a redundant packet without waiting for the feedback to arrive at the transmitter side, which enables the receiver to possibly decode $\mathbf{m}_{k'}$ without waiting for the next Stop-and-Wait HARQ round.

In order to keep the same energy at each time-slot, the superposed packet, belonging to the second layer, uses $100(1-\alpha)\%$ of the predefined energy per time-slot, while the packet sent by the first layer uses $100\alpha\%$ of the energy, with $\alpha \in [0,1]$. The influence of α will be investigated in Section V. The transmit vector $\mathbf{x}_t$ is given by:

$$\begin{cases} \mathbf{p}_k(\ell), & \text{if no superposition,} \\ \sqrt{\alpha}\mathbf{p}_k(\ell) + \sqrt{1-\alpha}\mathbf{p}_{k'}(\ell'), & \text{if superposition.} \end{cases}$$

We note that the case of $\alpha = 1$ corresponds to the conventional parallel Stop-and-Wait HARQ.

At the beginning of time-slot t the transmitter knows the ACK/NACK related to the messages sent up to time-slot $t-T$ (because of the feedback delay). According to this knowledge, the transmitter selects the packets to include in $\mathbf{x}_t$. As anticipated, the choice of the packet in the first layer corresponds to conventional parallel HARQ. Therefore, if packet $\mathbf{p}_k(\ell)$ was sent at time-slot $t-T$, the reception of a NACK relative to message $\mathbf{m}_k$ just before time-slot t triggers the transmission of another redundancy packet $\mathbf{p}_k(\ell+1)$, as long as $\ell < C$. Otherwise, the reception of an ACK of $\mathbf{m}_k$ triggers the transmission of a packet $\mathbf{p}_{k''}(1)$ associated with a new message $\mathbf{m}_{k''}$ (never transmitted before). The selection of the superposed packet in the second layer is done according to the following principles: *i)* superposing packets related to the most recent messages of the first layer to reduce the delay, *ii)* superposing unsent redundant packets to reduce the message error by using transmit diversity. Based on these principles, we describe the selection strategy by the following rules (ordered by priority), which determine the choice of the superposed packet in the second layer:

1) A packet $\mathbf{p}_{k'}(\ell')$ cannot be superposed if message $\mathbf{m}_{k'}$ is in timeout or previously ACKed.
2) As long as there are unacknowledged messages with unsent packets, the superposed packet is the unsent packet of the lowest index ℓ' of the most recent message $\mathbf{m}_{k'}$, with $k' \neq k$ (different messages in the two layers).
3) If the transmitter already sent all the packets of all the unacknowledged messages that are not in timeout, the superposed packet is the packet with the lowest index ℓ' that was not previously sent in the second layer. (Notice that this packet has been already sent once, in the first layer).
4) No packet is superposed to a packet of the first layer that has $\ell = C$.

The first rule prevents larger delays than those provided by conventional parallel HARQ, while the fourth rule reduces the probability to drop messages by forbidding interference during the last retransmission. According

time-slot	1	2	3	4	5	6	7	8	9
layer 1	$\mathbf{p}_1(1)$	$\mathbf{p}_2(1)$	$\mathbf{p}_3(1)$	$\mathbf{p}_1(2)$	$\mathbf{p}_2(2)$	$\mathbf{p}_4(1)$	$\mathbf{p}_1(3)$	$\mathbf{p}_5(1)$	$\mathbf{p}_4(2)$
layer 2		$\mathbf{p}_1(2)$	$\mathbf{p}_2(2)$	$\mathbf{p}_3(2)$	$\mathbf{p}_3(3)$	$\mathbf{p}_1(3)$		$\mathbf{p}_4(2)$	$\mathbf{p}_5(2)$
$\mathcal{F}_t$	$\{1\}_N$	$\{1,2\}_N$	$\{2,3\}_A, \{1\}_N$	$\{1\}_N$	$\{1\}_N$	$\{1,4\}_N$	$\{1,4\}_N$	$\{4,5\}_N$	$\{4,5\}_N$

TABLE I: A realization of the proposed protocol.

to this protocol, one can check that, at each time-slot, at most T messages are not previously ACKed nor in timeout, which means that the feedback at each time-slot contains at most T ACK/NACKs.

In Table I, we provide one example of our protocol, with $C = 3$ and $T = 3$. We denote by $\{\cdot\}_A$ (resp. $\{\cdot\}_N$) the set of message indexes triggering ACK (resp. NACK) feedback. The notation $\mathcal{F}_t$ stands for the output of the receiver at time-slot t. We remind that we consider instantaneous decoding at the end of the time-slot t, but $\mathcal{F}_t$ will be available at the transmitter side as a delayed feedback after T time-slots.

IV. Receiver Analysis

At the end of time-slot t, the receiver considers the observations of the most recent CT time-slots, corresponding to the maximum delay of the conventional parallel Stop-and-Wait HARQ. Since there are T parallel HARQ processes, there are at most T undecoded messages in this observation window. Therefore, the receiver attempts to decode these messages.

The output of the receiver is the feedback vector $\mathcal{F}_t$, which will be available at the transmitter at the beginning of time-slot $t+T$. The feedback vector $\mathcal{F}_t$ contains the ACK/NACK bits corresponding to the messages that *i)* are object of decoding at time-slot t, and *ii)* will not be in timeout at time-slot $t+T$. We notice that attempting to decode all the messages, including the ones that will be in timeout, is beneficial because it helps in removing the interference introduced by the superposition. This can be seen in the example in Table I. In time-slot 8, the receiver attempts to decode $\mathbf{m}_1$, $\mathbf{m}_4$ and $\mathbf{m}_5$. Since $\mathbf{m}_1$ will be in timeout by time-slot 10, $\mathcal{F}_8$ contains only information about $\mathbf{m}_4$ and $\mathbf{m}_5$. At time-slot 11, when $\mathcal{F}_8$ will be available at the transmitter, any feedback information about $\mathbf{m}_1$ would be useless. However, attempting to decode $\mathbf{m}_1$ is beneficial since it allows to remove the interference with message $\mathbf{m}_4$ on time-slot 6.

In the next Subsection we give an information theoretic characterization of the performance of the receiver.

Information theoretic characterization of the receiver

Let $\mathcal{M}$ be the set of messages that the receiver is attempting to decode at time-slot t. If the receiver successfully decodes the subset $\mathcal{D} \subseteq \mathcal{M}$ and none of the messages in $\mathcal{M} \setminus \mathcal{D}$, we say that the decoder operates in the rate region $\mathcal{R}_\mathcal{D}$. The set $\mathcal{D}$, along with the rules of the transmit protocol, allows to obtain $\mathcal{F}_t$. In order to characterize the decoding outcome, we *i)* evaluate the rate region $\mathcal{R}_\mathcal{D}$ for every possible $\mathcal{D} \subseteq \mathcal{M}$; and *ii)* determine, on the basis of the available observations, the operating rate region $\mathcal{R}_\mathcal{D}$ of the receiver. By definition, $\mathcal{R}_\mathcal{D}$ is given by the union of rate regions where the messages in $\mathcal{D}$ are successfully decoded (alone or jointly with other messages in $\mathcal{M} \setminus \mathcal{D}$), excluding the regions where the messages in $\mathcal{D}$ are jointly decoded with at least another message in $\mathcal{M} \setminus \mathcal{D}$. By construction of the system, the receiver can see the messages as users of a MAC channel. For a set of users $\mathcal{S}$, $\mathscr{R}_{MAC(\mathcal{S})}$ is the MAC rate region of users $\mathcal{S}$ considering the messages from users outside $\mathcal{S}$ as noise [12]. We consider first the case $\mathcal{D} \neq \emptyset$. The region where the messages in $\mathcal{D}$, and possibly other messages in $\mathcal{M}$, are successfully decoded is the union of the MAC rate regions of any set of users that includes $\mathcal{D}$, i.e., $\bigcup_{\mathcal{D}\subseteq\mathcal{S}} \mathscr{R}_{MAC(\mathcal{S})}$ [12]. The region where the messages in $\mathcal{D}$ are successfully decoded, jointly with at least another message in $\mathcal{M}$, is the union of the MAC regions of any set that includes $\mathcal{D}$ and at least another user from $\mathcal{M} \setminus \mathcal{D}$, i.e., $\bigcup_{\mathcal{D}\subset\mathcal{S},\mathcal{S}\neq\mathcal{D}} \mathscr{R}_{MAC(\mathcal{S})}$ [12]. We deduce the rate region in (2):

$$\begin{aligned}\mathcal{R}_\mathcal{D} &= \left(\bigcup_{\mathcal{D}\subseteq\mathcal{S}} \mathscr{R}_{MAC(\mathcal{S})}\right) \bigcap \left(\overline{\bigcup_{\substack{\mathcal{D}\subset\mathcal{S}\prime,\\ \mathcal{S}\prime\neq\mathcal{D}}} \mathscr{R}_{MAC(\mathcal{S}\prime)}}\right) \\ &= \mathscr{R}_{MAC(\mathcal{D})} \bigcap \left(\bigcap_{\mathcal{D}\subset\mathcal{S},\mathcal{S}\neq\mathcal{D}} \overline{\mathscr{R}_{MAC(\mathcal{S})}}\right). \quad (2)\end{aligned}$$

Since the regions $\mathcal{R}_\mathcal{D}$, for all possible $\mathcal{D} \subseteq \mathcal{M}$ form a partition by construction, the region $\mathcal{R}_{\mathcal{D}=\emptyset}$ is the complementary of the union of all rate regions for $\mathcal{D} \neq \emptyset$, i.e.,

$$\mathcal{R}_\emptyset = \overline{\bigcup_{\mathcal{D}\subseteq\mathcal{M},\mathcal{D}\neq\emptyset} \mathcal{R}_\mathcal{D}} = \bigcap_{\mathcal{D}\subseteq\mathcal{M},\mathcal{D}\neq\emptyset} \overline{\mathcal{R}_\mathcal{D}}. \quad (3)$$

Then, to determine whether the receiver operates in $\mathcal{R}_\mathcal{D}$, for any $\mathcal{D} \subseteq \mathcal{M}$, it is enough to verify whether the receiver operates within or outside the set of MAC regions involved in (2) and (3). The receiver operates in the MAC rate region $\mathscr{R}_{MAC(\mathcal{S})}$, for a set of messages $\mathcal{S}$, if the following set of inequalities is satisfied [12]:

$$\sum_{j\in\mathcal{T}} R_j \leq I(X_\mathcal{T}; Y | X_{\mathcal{S}\setminus\mathcal{T}}), \text{ for all } \mathcal{T} \subseteq \mathcal{S}, \quad (4)$$

where Y represents the observations in the window of size CT time-slots, $X_\mathcal{T}$ represents the sent packets relative to the messages in $\mathcal{T}$, and $X_{\mathcal{S}\setminus\mathcal{T}}$ is interpreted

likewise. The packets relative to messages that are not in $\mathcal{S}$ but are in Y are treated as interference. We also have $R_j = R$. The mutual information $I(X_{\mathcal{T}}; Y|X_{\mathcal{S}\setminus\mathcal{T}})$ can be calculated by reading the observations in the window of size CT, and cumulating the mutual information corresponding to the messages in $\mathcal{T}$. In this process, we need to consider that: 1) some packets are superposed, and sent with different power fractions, 2) the same packet may be transmitted more than once, 3) messages which have been already decoded in the past may allow to eliminate interfering packets in the observations. In Table I, $\mathcal{R}_{\mathcal{D}}$ corresponding to $\mathcal{D} = \{\mathbf{m}_2, \mathbf{m}_3\}$ at time-slot $t = 3$ is obtained thanks to (2) as:

$$\mathcal{R}_{\mathcal{D}} = \mathscr{R}_{MAC(\{\mathbf{m}_2,\mathbf{m}_3\})} \bigcap \overline{\mathscr{R}_{MAC(\{\mathbf{m}_1,\mathbf{m}_2,\mathbf{m}_3\})}}. \quad (5)$$

The MAC rate region $\mathscr{R}_{MAC(\{\mathbf{m}_2,\mathbf{m}_3\})}$ is given by:

$$\begin{cases} R \leq \log(1 + \frac{\alpha g_2}{1+(1-\alpha)g_2}) + \log(1 + (1-\alpha)g_3); \\ R \leq \log(1 + \alpha g_3); \\ 2R \leq \log(1 + \frac{\alpha g_2}{1+(1-\alpha)g_2}) + \log(1 + g_3), \end{cases}$$

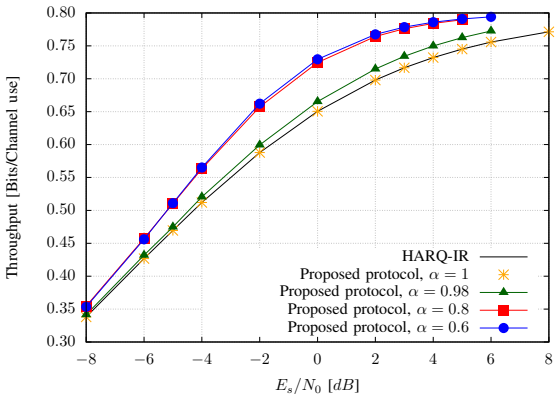
where packets corresponding to message $\mathbf{m}_1$ are considered as interference. $\mathscr{R}_{MAC(\{\mathbf{m}_1,\mathbf{m}_2,\mathbf{m}_3\})}$ is given by:

$$\begin{cases} R \leq \log(1 + g_1) + \log(1 + (1-\alpha)g_2); \\ R \leq \log(1 + \alpha g_2) + \log(1 + (1-\alpha)g_3); \\ R \leq \log(1 + \alpha g_3); \\ 2R \leq \log(1 + g_1) + \log(1 + g_2) + \log(1 + (1-\alpha)g_3); \\ 2R \leq \log(1 + g_1) + \log(1 + (1-\alpha)g_2) + \log(1 + g_3); \\ 2R \leq \log(1 + \frac{\alpha g_2}{1+(1-\alpha)g_2}) + \log(1 + g_3); \\ 3R \leq \log(1 + g_1) + \log(1 + g_2) + \log(1 + g_3). \end{cases}$$

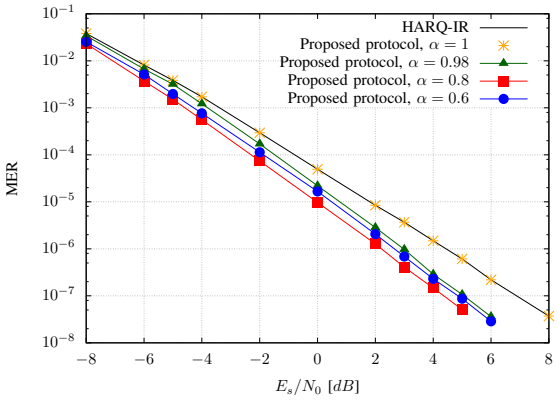
This characterization provides the receiver's performance for capacity-achieving codes.

V. Numerical Results

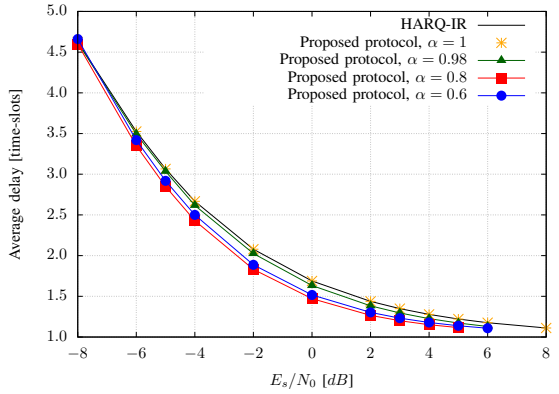
We present here numerical results, via computer simulations, of the proposed protocol in comparison to conventional parallel Stop-and-Wait HARQ, both with $C = 4$, $T = 3$ and $R = 0.8$, for capacity-achieving codes. HARQ-IR is implemented as described in Section II. In Fig. 1a, we plot the throughput which is the average number of correctly received information bits per channel use. The proposed protocol offers significant throughput gain in comparison to conventional parallel HARQ for any Signal to Noise Ratio (SNR) at $\alpha = 0.6$ and $\alpha = 0.8$. E_s is the energy consumed for sending one symbol. In case of superposition, we remind that the energy is shared between superposed symbols with the proportion α for the layer 1. The proposed protocol also achieves lower Message Error Rate (MER) than the conventional parallel HARQ, as it can be seen in Fig. 1b. The MER is defined as the average ratio of the number of dropped messages over the number of sent messages. The performance of the proposed protocol depends on



(a) Throughput of the proposed protocol.



(b) MER of the proposed protocol.



(c) Average delay of the proposed protocol.

Fig. 1: Performance of the proposed protocol.

the choice of the power fraction α. Therefore, we plot in Fig. 2 both throughput and MER, at $E_s/N_0 = 0dB$, versus α. The power fraction α can be numerically optimized and fixed for each desired SNR depending on the application requirements. Further optimization of the power allocation is possible, but is out of the scope of this work. The average delay, which is the average number of elapsed time-slots until the receiver successfully decodes a message, is presented in Fig. 1c.

In addition to lower average delay, the proposed

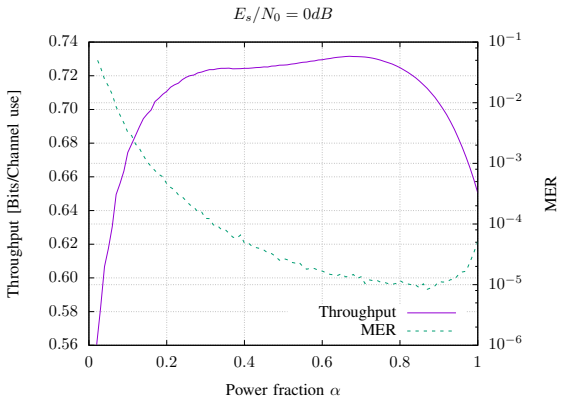
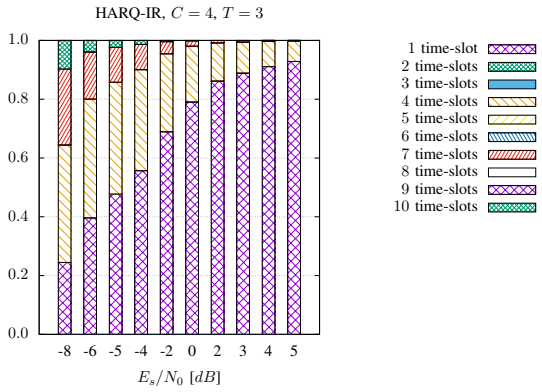
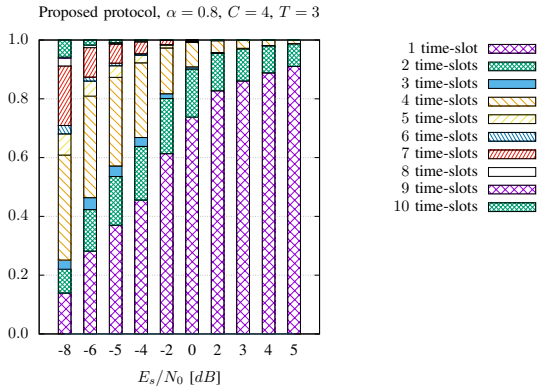


Fig. 2: Throughput and MER of the proposed protocol for different power fractions α at $0dB$.



(a) Conventional parallel HARQ.



(b) Proposed protocol.

Fig. 3: Comparison of the delay distributions.

protocol offers a better delay distribution. The delay distribution represents the proportion of successfully delivered messages for each value of delay. Fig. 3a shows that, due to parallel Stop-and-Wait HARQ protocol, retransmissions occur every $T = 3$ time-slots. Thus, a message can be decoded only with a delay of 1 (by decoding the first packet), 4 (by decoding the first retransmission), 7 (by decoding the second retransmission) or 10 time-slots (by decoding the last retransmission) for $C = 4$. However, due to superposition in the proposed protocol, the receiver can decode a message with a finer granularity of delays, and delays of 1, 2,...,10 time-slots are possible, as it can be seen in Fig. 3b. We observe that the probability to have higher delay (such as 4 or more) is smaller with our proposed protocol.

VI. Conclusion

A multi-packet HARQ protocol when feedback is delayed has been proposed. Information-theoretic analysis of the receiver shows that the proposed protocol offers a smaller average delay, a better delay distribution, a higher throughput, and a lower message error rate than the conventional parallel Stop-and-Wait HARQ. The proposed protocol requires at the receiver side a joint decoding of the superposed packets. Future works might focus on implementing the protocol with practical channel encoders and decoders, creating other new protocols with respect to other rules depending on the application, or managing more relevantly the interference at the transmitter side since the messages are sent by the same transmitter.

References

[1] S. Sesia, I. Toufik, and M. Baker, "LTE: the Long Term Evolution - from Theory to practice," Wiley, 2009.

[2] S. Lin and D. Costello, "Error Control Coding," Pearson, 2005.

[3] A. E. Hamss, L. Szczecinski and P. Piantanida, "Increasing the throughput of HARQ via multi-packet transmission," IEEE Global Communications Conference , Austin, TX, 2014.

[4] M. Jabi, A. E. Hamss, L. Szczecinski and P. Piantanida, "Multi-packet Hybrid ARQ: Closing Gap to the Ergodic Capacity," IEEE Transactions on Communications, vol. 63, no. 12, pp. 5191-5205, Dec. 2015.

[5] R. Zhang and L. Hanzo, "Superposition-Coding-Aided Multiplexed Hybrid ARQ Scheme for Improved End-to-End Transmission Efficiency," IEEE Transactions on Vehicular Technology, vol. 58, no. 8, pp. 4681-4686, Oct. 2009.

[6] F. Takahashi and K. Higuchi, "HARQ for Predetermined-Rate Multicast Channel," IEEE Vehicular Technology Conference, Taipei, Taiwan, 2010..

[7] A. N. Assimi, C. Poulliat and I. Fijalkow, "Packet combining for multi-layer hybrid-ARQ over frequency-selective fading channels," European Signal Processing Conference, Glasgow, 2009.

[8] A. Steiner and S. Shamai, "Multi-layer broadcasting hybrid-ARQ strategies for block fading channels," IEEE Transactions on Wireless Communications, vol. 7, no. 7, pp. 2640-2650, July 2008.

[9] K. F. Trillingsgaard and P. Popovski, "Generalized HARQ Protocols with Delayed Channel State Information and Average Latency Constraints," IEEE Transactions on Information Theory, vol. PP, no. 99, pp. 1-1, 2018.

[10] L. Szczecinski, S. R. Khosravirad, P. Duhamel and M. Rahman, "Rate Allocation and Adaptation for Incremental Redundancy Truncated HARQ," IEEE Transactions on Communications, vol. 61, no. 6, pp. 2580-2590, June 2013.

[11] T. Villa, R. Merz, R. Knopp and U. Takyar, "Adaptive modulation and coding with hybrid-ARQ for latency-constrained networks," European Wireless Conference, Poznan, Poland, 2012.

[12] B. Bandemer, A. E. Gamal and Y. H. Kim, "Simultaneous nonunique decoding is rate-optimal," Annual Allerton Conference on Communication, Control, and Computing, Monticello, IL, 2012.