Low-Latency Short-Packet Transmissions: Fixed Length or HARQ?

Downloaded from: https://research.chalmers.se, 2019-10-27 12:04 UTC

Citation for the original published paper (version of record):
Low-Latency Short-Packet Transmissions: Fixed Length or HARQ?
[Source Title missing]
http://dx.doi.org/10.1109/GLOCOMW.2018.8644397

N.B. When citing this work, cite the original published paper.
Low-Latency Short-Packet Transmissions: Fixed Length or HARQ?

Johan Östman, Rahul Devassy, Guido Carlo Ferrante, Giuseppe Durisi
Chalmers University of Technology, Gothenburg, Sweden

Abstract—We study short-packet communications, subject to latency and reliability constraints, under the premises of limited frequency diversity and no time diversity. The question addressed is whether, and when, hybrid automatic repeat request (HARQ) outperforms fixed-blocklength schemes with no feedback (FBL-NF) in such a setting. We derive an achievability bound for HARQ, under the assumption of a limited number of transmissions. The bound relies on pilot-assisted transmission to estimate the fading channel and scaled nearest-neighbor decoding at the receiver. We compare our achievability bound for HARQ to state-of-the-art achievability bounds for FBL-NF communications and show that for a given latency, reliability, number of information bits, and number of diversity branches, HARQ may significantly outperform FBL-NF. For example, for an average latency of 1 ms, a target error probability of $10^{-3}$, 30 information bits, and 3 diversity branches, the gain in energy per bit is about 4 dB.

I. INTRODUCTION

Fifth generation (5G) wireless systems are envisioned to support communications with stringent requirements on reliability and latency—the so-called ultra-reliable low-latency communications (URLLC) [1]. The objective is to enable use cases like smart grids and wireless industrial control. While previous-generation wireless systems focused mostly on the support of large data rates, URLLC systems target transmission of small data payloads, carried by short coded packets to meet the latency requirements.

Performance analyses of wireless systems are often carried out using asymptotic information-theoretic metrics like ergodic capacity and outage capacity, under the implicit assumption that a message is encoded into an arbitrarily large number of coded symbols. Although a good approximation for systems where latency concerns are secondary, such performance analyses are not suitable for URLLC [2]. Instead, to study the performance of URLLC links, one must resort to tools that allow for the analysis of communication systems in the nonasymptotic regime where a message is encoded onto a finite, often small, number of symbols.

For communications with fixed blocklength and no feedback (FBL-NF), the nonasymptotic information-theoretic tools developed in [3] have recently enabled the characterization of the maximum coding rate achievable for a given blocklength and error probability in many scenarios of practical relevance for 5G and beyond. For example, it has allowed for the study of the rate achievable with short packets on general quasi-static fading channels [4] and multiple-input multiple-output (MIMO) Rayleigh block-fading channels [5]. Furthermore, the practically relevant case of pilot-assisted transmission (PAT) and scaled nearest-neighbor (SNN) decoding has been analyzed for single-input single-output (SISO) Rician block-fading channels [6] and for MIMO Rayleigh block-fading channels [7].

Finite-blocklength information theory prescribes that for several channels, including the AWGN channel, short-packet transmission incurs a rate loss from capacity roughly proportional to the inverse of the square root of the blocklength [3]. However, when one is allowed to use variable-length stop-feedback (VLSF) codes, the picture is different. Specifically, capacity is approached much faster in the blocklength [8]. VLSF codes are an instance of variable-length feedback codes where the encoder keeps on transmitting coded symbols until it receives a single-bit feedback sent by the decoder to inform the encoder that decoding is complete. Incremental redundancy hybrid automatic repeat request (HARQ) is an instance of VLSF codes. In [8], decoding is attempted upon reception of every new symbol. This setup was later extended to blocks of symbols in [9], a scenario that is more relevant for HARQ. Under the assumption that the received symbols are discarded and transmission starts over if decoding is not successful after a finite number of blocks, the rate penalty resulting from decoding after each block of symbols rather than after each symbol was characterized in [9] for the AWGN channel. However, such strategy still outperforms FBL-NF codes for moderate latencies.

In [10], FBL-NF and HARQ are compared for the AWGN channel by taking into account the delay caused by the transmission of a positive/negative acknowledgment (ACK/NACK). The analysis, however, does not account for the event of an undetected error, i.e., the event that an ACK is fed back although the decision of the decoder is wrong. Furthermore, the normal approximation [3] is used in the analysis. However, this approximation, which is based on the Berry-Esseen theorem, is accurate only when the blocklength and the error probability are not too small and may not be suitable for URLLC.

Contributions: We consider an HARQ scheme employing PAT and SNN decoding at the receiver for transmission of short data packets over a SISO block-fading channel. Leveraging [8, Thm. 3], which provides an extension of the dependence-testing bound [3, Thm. 17] for FBL-NF to VLSF codes, we derive an achievable bound on the minimum energy per bit required to transmit a small information payload under a given latency and reliability target. Differently from [8, Thm. 3], where decoding stops when the accumulated information density corresponding to one of the codewords exceeds a threshold, in our setup decoding stops when the accumulated generalized information density exceeds a threshold or the number of transmissions reaches a maximum predetermined value.

Our bound depends on the cumulative distribution function

This work was supported by the Swedish Research Council under grants 2014-6066 and 2016-03293.
We consider a memoryless SISO Rayleigh block-fading channel in which fading is assumed to stay constant within a coherence block and to change independently across coherence blocks. The channel coherence time and coherence bandwidth are denoted by $T_c$ and $B_c$, respectively. For a system bandwidth $B$, the available number of diversity branches is $L_c = \lceil B/B_c \rceil$. The duration of a transmitted codeword is assumed to be much smaller than the coherence time. Hence, no time diversity is available. This is in line with commonly used fourth generation (4G) and 5G channel models, as discussed in Section IV. We consider orthogonal frequency-division multiplexing (OFDM) and assume that a resource block (RB) consists of $d$ OFDM symbols, each spanning $u$ subcarriers. Therefore, an RB consists of $n_c = ud$ symbols. The time interval over which an RB is transmitted is referred to as slot, in accordance with the 3rd generation partnership project (3GPP) terminology. We let $L$ be the number of diversity branches used in a slot out of the available $L_c$. Hence, $L$ RBs, amounting to $Ln_c$ symbols, are transmitted in each slot, as shown in Fig. 1.

We shall consider two communication schemes: FBL-NF and HARQ. In the FBL-NF case, a codeword is transmitted over $L$ diversity branches per slot and $v$ slots in total. Hence, a codeword consists of $vLn_c$ symbols, see Fig. 1. We assume that transmissions on different slots occur on different diversity branches, so that transmission symbols in different slots experience independent fading. This means that the number of slots per codeword is no larger than $\ell_{\text{max}} = \lfloor L_c/L \rfloor$. At the receiver side, decoding is performed once the entire codeword is received.

In the HARQ case, we consider one direction of a bidirectional communication link. We assume that communication is organised in rounds, each consisting of two consecutive slots within the same diversity branches (time-division duplexing). The two communicating devices are assigned one slot each in every round. We assume that transmissions on different rounds occur on different diversity branches (see Fig. 2). Hence, transmitted symbols in different rounds experience independent fading. Consequently, the maximum number of rounds allowed is $\ell_{\text{max}}$. In each round, a device transmits $Ln_c$ symbols during one slot and listens for an ACK/NACK and possibly other data during the next slot. At the other device (receiver), decoding is attempted after the first slot. In the second slot, the receiver piggybacks an ACK/NACK on its data, to inform the transmitter whether to continue or to terminate transmission. If a NACK is piggybacked, incremental redundant symbols are sent during the next round. Rounds go on until the transmitter observes an ACK or it reaches the maximum predetermined number of rounds. Such an ACK/NACK transmission introduces a delay compared to FBL-NF transmission, which we will take into account in our analysis. We assume that the feedback delay per round amounts to one slot. However, our analysis can be easily generalized to arbitrary feedback delays. As illustrated in Fig. 2, when the same amount of resources is used in HARQ (per communication direction) and FBL-NF, the FBL-NF scheme requires half of the time of the HARQ scheme.

**System Model**

### Setup

We consider orthogonal frequency-division multiplexing (OFDM) and assume that a resource block (RB) consists of $d$ OFDM symbols, each spanning $u$ subcarriers. Therefore, an RB consists of $n_c = ud$ symbols. The time interval over which an RB is transmitted is referred to as slot, in accordance with the 3rd generation partnership project (3GPP) terminology. We let $L$ be the number of diversity branches used in a slot out of the available $L_c$. Hence, $L$ RBs, amounting to $Ln_c$ symbols, are transmitted in each slot, as shown in Fig. 1.

We shall consider two communication schemes: FBL-NF and HARQ. In the FBL-NF case, a codeword is transmitted over $L$ diversity branches per slot and $v$ slots in total. Hence, a codeword consists of $vLn_c$ symbols, see Fig. 1. We assume that transmissions on different slots occur on different diversity branches, so that transmission symbols in different slots experience independent fading. This means that the number of slots per codeword is no larger than $\ell_{\text{max}} = \lfloor L_c/L \rfloor$. At the receiver side, decoding is performed once the entire codeword is received.

In the HARQ case, we consider one direction of a bidirectional communication link. We assume that communication is organised in rounds, each consisting of two consecutive slots within the same diversity branches (time-division duplexing). The two communicating devices are assigned one slot each in every round. We assume that transmissions on different rounds occur on different diversity branches (see Fig. 2). Hence, transmitted symbols in different rounds experience independent fading. Consequently, the maximum number of rounds allowed is $\ell_{\text{max}}$. In each round, a device transmits $Ln_c$ symbols during one slot and listens for an ACK/NACK and possibly other data during the next slot. At the other device (receiver), decoding is attempted after the first slot. In the second slot, the receiver piggybacks an ACK/NACK on its data, to inform the transmitter whether to continue or to terminate transmission. If a NACK is piggybacked, incremental redundant symbols are sent during the next round. Rounds go on until the transmitter observes an ACK or it reaches the maximum predetermined number of rounds. Such an ACK/NACK transmission introduces a delay compared to FBL-NF transmission, which we will take into account in our analysis. We assume that the feedback delay per round amounts to one slot. However, our analysis can be easily generalized to arbitrary feedback delays. As illustrated in Fig. 2, when the same amount of resources is used in HARQ (per communication direction) and FBL-NF, the FBL-NF scheme requires half of the time of the HARQ scheme.
The input-output relation for the slot assigned for the forward transmission in round $j$, satisfies

$$P_{Y_{j}} = \prod_{k=1}^{L} P_{Y_{j,k}} | X_{j,k} (1)$$

with

$$Y_{j,k} = H_{j,k} X_{j,k} + W_{j,k}. \quad (2)$$

Here, $Y_{j} = [Y_{j,1}, \ldots, Y_{j,L}]$ and $X_{j} = [X_{j,1}, \ldots, X_{j,L}]$, where $Y_{j,k} \in \mathbb{C}^{n_c}$ and $X_{j,k} \in \mathbb{X} = \{ x \in \mathbb{C}^{n_c} : ||x||^2 = n_c \rho \}$ for $j = 1, 2, \ldots, L$. The variable $\rho$ denotes the SNR. The Rayleigh fading is modeled by $H_{j,k} \sim \mathcal{CN}(0, 1)$, and $W_{j,k} \sim \mathcal{CN}(0, I_{n_c})$ is the AWGN noise. The random variables $\{H_{j,k}\}$ and $\{W_{j,k}\}$, which are mutually independent, are also independent over $j$ and $k$. No a priori knowledge of the realizations of $\{H_{j,k}\}$ is available at the transmitter and at the receiver. Consequently, it is reasonable to transmit equal-power signals over the available diversity branches. This justifies our assumption that $X_{j,k} \in \mathbb{X}$ for all $j$ and $k$.

A code for the channel (1) in the HARQ setup is formally defined next by adapting the notion of VLSF codes in [8].

**Definition 1:** An $(\ell, M, \epsilon, \rho, \ell_{\text{max}})$-VLSF code, where $\ell \geq 1$, $M$, $\ell_{\text{max}}$ are positive integers, $\rho > 0$, and $0 < \epsilon < 1$, consists of

1. A random variable $U$ with distribution $P_{U}$ defined on a set $\mathcal{U}$ with $|\mathcal{U}| \leq 2$ that is revealed to both the transmitter and the receiver before the start of transmission. $U$ acts as a common randomness and enables the use of randomized encoding and decoding strategies.
2. An encoder $f : \mathcal{U} \times \{1, \ldots, M\} \rightarrow \mathcal{X}^{L_{\text{max}}}$, that maps a message $J$, which is uniformly distributed on $\{1, \ldots, M\}$, to a codeword in the set $\{c(1), \ldots, c(M)\}$. Each codeword is structured as $c(m) = [c_{1}(m), \ldots, c_{\ell_{\text{max}}}(m)]$ where $c_{j}(m) \in \mathcal{X}$ for $j = 1, \ell_{\text{max}}$, and $m = 1, 1, \ldots, M$.
3. A sequence of decoders $g_{v} : \mathcal{U} \times \mathbb{C}^{n_{c} \cdot L_{v}} \rightarrow \{1, \ldots, M\}$, $1 \leq v \leq \ell_{\text{max}}$, and a stopping time $\tau^{*}$, that is adapted to the filtration $\{ \sigma(U, Y_{1}, \ldots, Y_{v}) \}_{v=1}^{\ell_{\text{max}}}$, and satisfies both

$$\mathbb{E}[\tau^{*}] \leq \ell, \quad (3)$$

and the average packet error probability target

$$\mathbb{P} \{ g_{\tau^{*}}(U, Y_{1}, \ldots, Y_{v}) \neq J \} \leq \epsilon. \quad (4)$$

After the stopping time is triggered, the decoder uses the feedback channel to inform the encoder, through a one-bit ACK, to stop the transmission of the current message and to move to the transmission of the next one. We assume throughout the paper that this bit is error free and that so are the NACK bits transmitted in the previous feedback rounds. Differently from most literature, we will consider in our analysis undetected error events in which an ACK is fed back although the decoder has chosen the wrong message.

For a given $\ell$, $\epsilon$, and $\ell_{\text{max}}$, the maximum coding rate $R^{*}$, measured in information bits per channel use is defined as

$$R^{*}(\ell, \epsilon, \rho, \ell_{\text{max}}) = \sup \left\{ \frac{\log_{2} M}{\ell L n_{c}} : \exists (\ell, M, \epsilon, \rho, \ell_{\text{max}}) \right\} \quad \text{- VLSF code}. \quad (5)$$

We will also be interested in the problem of minimizing the average number of transmissions $\ell$ for a given number of messages $M$, which yields the following definition:

$$\ell^{*}(\epsilon, \rho, \ell_{\text{max}}) = \inf \{ \ell : \exists (\ell, M, \epsilon, \rho, \ell_{\text{max}}) \} \quad \text{- VLSF code}. \quad (6)$$

Some of our results will also be expressed in terms of the minimum energy per bit$^{1}$ $E_{b}^{*}/N_{0}$, which is related to $\ell^{*}$ as

$$\frac{E_{b}^{*}}{N_{0}} = \frac{\rho \log_{2} M}{\ell^{*} L n_{c}} \quad \text{VLSF code}. \quad (7)$$

The corresponding metrics for FBL-NF are defined as in [6].

**B. PAT with SNN Decoding**

Following [6], we assume that, for slot $j$ and coherence interval $k$, the input vector $X_{j,k}$ is of the form $[x^{(b)}, X_{j,k}^{(d)}]$, where $x^{(b)} \in \mathbb{C}^{n_{b}}$, $1 \leq n_{b} < n_{c}$, is a deterministic vector containing pilot symbols and $\|x^{(b)}\|^2 = n_{p} \rho$. The vector $X_{j,k}^{(d)} \in \mathbb{C}^{n_{d}}$ contains the $n_{d} = n_{c} - n_{b} - n_{p}$ data symbols. Let $P_{X^{(d)}}$ denote a distribution on $\mathbb{C}^{n_{d}}$ such that $X_{j,k} \in \mathcal{X}$ w.p.1 when $X_{j,k}^{(d)} \sim P_{X^{(d)}}$. Let $Y_{j,k}^{(b)}$ and $Y_{j,k}^{(d)}$ be the vectors containing the received samples corresponding to the pilot and data symbols, respectively. Given $Y_{j,k}^{(b)}$ and $Y_{j,k}^{(d)}$, the receiver computes the maximum likelihood estimate $\hat{H}_{j,k}$ of the fading coefficient $H_{j,k}$ as

$$\hat{H}_{j,k} = \frac{1}{n_{p} \rho} (x^{(p)})^{H} y^{(p)}_{j,k}. \quad (8)$$

Next we define the SNN decoder. Denote a candidate codeword and the observed channel outputs up to the $v$th slot by $x^{(v)} = [x_{1}, \ldots, x_{v}, L_{v}]$, and $y^{(v)} = [y_{1}, \ldots, y_{v}, L_{v}]$, respectively, where $x_{j,k} \in \mathcal{X}$ and $y_{j,k} \in \mathbb{C}^{n_{c}}$ for $j = 1, \ldots, v$, and $k = 1, \ldots, L_{v}$. Here, $X_{j,k}$ and $Y_{j,k}$ follow the structure outlined above. The SNN metric is given by

$$q^{(v)}(x^{(v)}, y^{(v)}) = \prod_{j=1}^{v} \prod_{k=1}^{L_{j}} q(x_{j,k}, y_{j,k}) \quad (9)$$

where

$$q(x_{j,k}, y_{j,k}) = \prod_{i=1}^{n_{d}} \exp \left( -|y_{j,k,i}^{(d)} - \hat{H}_{j,k} x_{j,k,i}^{(d)}|^{2} \right). \quad (10)$$

Here, $y_{j,k,i}$ and $x_{j,k,i}$ denote the $i$th entry of $y_{j,k}^{(d)}$ and $x_{j,k}^{(d)}$, respectively. We refer to this coding scheme as PAT-SNN.

Finally, we define the generalized information density in slot $v$ as a mapping $A^{v} \times \mathbb{C}^{n_{c} \cdot L_{v}} \rightarrow \mathbb{R}$, given as

$$r_{v}^{(s)}(x^{(v)}, y^{(v)}) = \sum_{j=1}^{v} \sum_{k=1}^{L_{j}} r_{v}(x_{j,k}, y_{j,k}) \quad (11)$$

where the generalized information density per coherence block $r_{s}(x_{j,k}, y_{j,k})$ is

$$r_{s}(x_{j,k}, y_{j,k}) = \log \frac{q(x_{j,k}, y_{j,k})^{s}}{\mathbb{E} \left[ q(x, y_{j,k})^{s} \right]}. \quad (12)$$

\footnote{We will not consider the energy sent to send the feedback bit reliably.}
Here, $\overline{X} = [x^{(p)}, X^{(d)}]$ with $x^{(p)}$ an arbitrary pilot vector satisfying the properties listed above, $X^{(d)} \sim P_{X^{(d)}}$, and $s \geq 0$.

For the special case in which $P_{X^{(d)}}$ is a product distribution, i.e., $P_{X^{(d)}}(x^{(d)}_{j,k}) = \prod_{i=1}^s P_X(x^{(d)}_{j,k,i})$, we can write (12) as

$$s_j(x_{j,k}, y_{j,k}) = \sum_{i=1}^s -s|y^{(d)}_{j,k,i} - \hat{h}_{j,k} x^{(d)}_{j,k,i}|^2$$

$$- \log \mathbb{E}_{\overline{X}} \left[ \exp \left( -s|y^{(d)}_{j,k,i} - \hat{h}_{j,k} x^{(d)}_j|^2 \right) \right]. \quad (13)$$

III. FINITE-BLOCKLENGTH ACHIEVABILITY BOUNDS

A. Fixed-Blocklength Transmission without Feedback

Next, we review an achievability bound for the channel in (2), based on the random-coding union bound with parameter $s$ [12, Thm. 1]. This bound will be used to assess the performance of PAT-SNN-FBL-NF transmission.

Theorem 1: Fix an integer $1 \leq n_d < n_c$, a rate $R$, and a real number $s \geq 0$. The average error probability for the PAT-SNN-NF-BF scheme, operating as described in Section II-B, is upper-bounded as

$$\epsilon \leq \mathbb{E} \left[ \exp \left( -s|x^{(s)}_j X^{(s)}, Y^{(s)} - (2^{sL \ln R} - 1)^{+} \right) \right] \quad (14)$$

where $X^{(s)} = [X_1, \ldots, X_1, X_{L,1}, \ldots, X_{L,L}]$ and each $X_{j,k} \sim P_{X^{(d)}}$, and $Y_{j,k}$ is the channel output according to (2), for $j = 1, \ldots, v$ and $k = 1, \ldots, L$.

Proof: See [6, Thm. 3].

Remark: The latency for the FBL-NF scheme is $T_{f} = v d T_{o}$, where $T_{o}$ is the OFDM symbol duration. In the HARQ case we will analyze next, latency is a random variable. We will investigate both its average value and its CDF.

B. HARQ Transmission

We provide next an achievability bound for the HARQ setup, which closely follows [8, Thm. 3]. To prove [8, Thm. 3], one computes the accumulated information density for each codeword and stops when one of the information densities exceeds a given threshold. In our case, instead of information density, we accumulate generalized information density, see (12). Furthermore, we only allow for a fixed number of transmission rounds, after which transmission is terminated and a decision is taken. This differs from the setup in [9] where, after a fixed number of transmission rounds, the received data is discarded and transmission is restarted. Incorporating these changes to [8, Thm. 3], one obtains the following.

Theorem 2 (Achievability HARQ): Fix three scalars $\gamma > 0$, $s \geq 0$, $\rho > 0$, and two positive integers $\ell_{\max}$ and $1 \leq n_p < n_c$. Let $(X_j^{(s)})_{j=1}^{\infty}$ be a stochastic process such that $X_j = [X_{j,1}, X_{j,2}, \ldots, X_{j,L}]$, and $X_{j,k} \in \mathcal{X}$ are identical and independently distributed (i.i.d.) for $j = 1, \ldots$ and $k = 1, \ldots, L$. Let $X_{j,k}^{(s)} = [x^{(p)}, X^{(d)}_{j,k}]$, where $x^{(p)}$ is an arbitrary pilot vector satisfying $||x^{(p)}||^2 = n_p \rho$ and $X^{(d)}_{j,k} \sim P_{X^{(d)}}$. Let $(X_j^{(s)})_{j=1}^{\infty}$ be an independent copy of $(X_j^{(s)})_{j=1}^{\infty}$. Moreover, define the stopping times

$$\tau = \inf \{ v \geq 1 : s|Y^{(s)}_v, Y^{(s)}_v|^{\gamma} \geq \gamma \}, \quad (15)$$

$$\ell = \min \{ \ell \leq \min \{ \ell_{\max}, \tau \} \}.$$
Remark: The average latency for the HARQ scheme is \( T_d = 2d_oT_m \), where the factor 2 is due to the feedback delay and \( T_m \) is the OFDM symbol duration. The maximum latency is, however, \( 2d_{\text{max}}T_m \). Note also that the CDF of the delay \( \tau \) is required to compute (25).

No closed-form expressions for (17) and (18) are available. Furthermore, evaluating (18) is numerically challenging. Indeed, \( M \) is typically very large (for a fixed rate \( R \), it grows exponentially in the blocklength), so the first probability is very small thus preventing the use of Monte Carlo methods. For the case of summands in (16) with negative drift, i.e., \( \mathbb{E}[t_s(\mathbf{X}_{k,1}, \mathbf{Y}_{k,1})] < 0 \), we can further relax (18) by upper-bounding \( P[\tau \leq \min\{\ell_{\text{max}}, \tau\}] \) using Wald’s identity as [14, Cor. 9.4.4]

\[
P[\tau \leq \min\{\ell_{\text{max}}, \tau\}] \leq P[\tau \leq \infty] \leq \exp(-\beta^*\gamma) \tag{26}
\]

where \( \beta^* \) is the positive solution of

\[
\kappa(\beta) = \log \mathbb{E}\left[ \exp\left( \beta \sum_{k=1}^{L} t_s(\mathbf{X}_{k,1}, \mathbf{Y}_{k,1}) \right) \right] = 0. \tag{27}
\]

In the upcoming section, we let \( P_{\mathbf{X}(o)} \) be a product distribution. Hence, the generalized information density is given by (13). It then follows from Jensen’s inequality that \( \mathbb{E}[t_s(\mathbf{X}_{k,1}, \mathbf{Y}_{k,1})] < 0 \). This enables us to use (26) in our numerical evaluations. Furthermore, we can express \( \kappa(\beta) \) as

\[
\kappa(\beta) = L \log \mathbb{E}\left[ \frac{\mathbb{E}[q(\mathbf{X}^{(d)}, \mathbf{Y}^{(d)}) | Y = y]}{\mathbb{E}[q(\mathbf{X}^{(d)}))]} \right] \tag{28}
\]

where \( \mathbf{X} = [\mathbf{x}^{(p)}, \mathbf{X}^{(d)}] \) with \( \mathbf{X}^{(d)} \sim P_{\mathbf{X}(o)} \) and \( \mathbf{Y} \) distributed as in (2), independent of \( \mathbf{X} \). From (28), we obtain \( \beta^* = 1 \).

Remark: An upper bound similar to (26) with \( \beta^* = 1 \) is obtained in [8, Eq. (113)] by using a change-of-measure argument. We cannot follow the same strategy here because of the mismatched decoding metric.

IV. NUMERICAL RESULTS

In this section, we compare the performance of HARQ and FBL-NF. For the sake of concreteness, we consider i.i.d. input symbols drawn uniformly from a quaternary phase-shift keying (QPSK) constellation. We extract our channel parameters from the TDL-C 300 ns–3 km/h channel model [15]. We choose the system bandwidth, subcarrier bandwidth, and OFDM symbol duration equal to 20 MHz, 15 kHz, and 71.4 \( \mu\text{s} \), respectively. This results in 30 available diversity branches. The channel model has coherence bandwidth \( B_c = 0.66 \text{ MHz} \), which implies \( u \leq 44 \). We let \( u = 24 \) and \( d = 3 \). Hence, an RB occupies 360 kHz and its duration is 214.2 \( \mu\text{s} \). We also let \( \epsilon = 10^{-3} \), which corresponds to the least stringent reliability constraint for URLLC [16]. The latency for HARQ is \( T_d = 2T_o\epsilon d \), and thus the minimum and maximum latency are \( 2T_o\epsilon d \) and \( 2T_o\epsilon d_{\text{max}} \), respectively. For FBL-NF, the minimum latency corresponds to one slot duration, i.e., \( T_o\epsilon d \), and the maximum latency is \( T_o\ell_{\text{max}} \). Throughout this section, we assume that each message contains \( k = \log_2 M = 30 \) information bits. The channel and system parameters are summarized in Table I.

Next, we detail how we evaluate the minimum energy per bit (7). The procedure for computing the maximum coding rate (5) is similar. In both FBL-NF and HARQ, we fix \( \rho, \epsilon, k, T_o, u, d, \) and \( L \). For HARQ, we find the smallest threshold \( \gamma \) such that the right-hand side of (18), relaxed using (26), is below the desired error target. This \( \gamma \) is then used to evaluate (17), which provides an upper bound on the average latency \( T_d \).

For FBL-NF, we search for the smallest \( \epsilon \) such that the right-hand side of (14) is below the desired error target. Throughout this section, all charts are obtained by optimizing over the parameter \( s \) and the number of pilot symbols \( n_p \).

In Fig. 3, we show the minimum energy per bit required to fulfill the reliability target as a function of average latency. It can be seen that, for fixed \( L \), HARQ significantly outperforms FBL-NF. For example, for \( L = 3 \) and \( T_d = 1 \text{ ms} \), the difference is about 4 dB. It can also be seen that \( E_o/N_0 \) does not decrease monotonically with the average latency for both HARQ and FBL-NF, but there exists a latency \( T_{d}^* \) for which \( E_o/N_0 \) is minimized. When \( T_d < T_{d}^* \), increasing the average blocklength provides an SNR gain that outweighs the rate penalty (see (7)), whereas the opposite is true when \( T_d > T_{d}^* \). HARQ performs better than FBL-NF because for a fixed \( \epsilon \), changes in the average blocklength result in larger changes in the required SNR than in FBL-NF. We also observe a tradeoff between the average latency and the number of diversity branches used: choosing a large \( L \) is more energy efficient for small average latency \( T_d \), since in this regime the number of possible retransmissions is small, and thus higher diversity is beneficial. As the average latency increases, choosing a smaller number of diversity branches is preferable.

In Fig. 4, we compare the maximum coding rate achievable with the two schemes as a function of the average latency. It can be seen that larger rates are achievable by using HARQ rather than FBL-NF. A similar behavior was reported for the AWGN channel in [9]. The rate increases with \( L \) in both systems, since a larger \( L \) allows for more diversity and longer codewords.

The results presented so far are in terms of average latency. However, average and maximum latencies are not equal in HARQ. Hence, one may argue that the above comparisons are unfair since the maximum latency with HARQ may be larger than the latency with FBL-NF. In Fig. 5, we illustrate the CDF of the latency for both FBL-NF and HARQ, for the case \( L = 2 \), and \( \rho \in [-5, -2] \) dB. The CDF of the FBL-NF latency is a

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( B_c )</td>
<td>50% coherence bandwidth</td>
<td>0.66 MHz</td>
</tr>
<tr>
<td>( T_o )</td>
<td>50% coherence time</td>
<td>85 ms</td>
</tr>
<tr>
<td>( L_o )</td>
<td>Number of available diversity branches</td>
<td>30</td>
</tr>
<tr>
<td>( k )</td>
<td>Information bits</td>
<td>30</td>
</tr>
<tr>
<td>( \epsilon )</td>
<td>Error probability</td>
<td>( 10^{-3} )</td>
</tr>
<tr>
<td>( B )</td>
<td>System bandwidth</td>
<td>20 MHz</td>
</tr>
<tr>
<td>( B_s )</td>
<td>Subcarrier bandwidth</td>
<td>15 kHz</td>
</tr>
<tr>
<td>( T_o )</td>
<td>OFDM symbol duration</td>
<td>71.4 ( \mu\text{s} )</td>
</tr>
<tr>
<td>( u )</td>
<td>Subcarriers per RB</td>
<td>24</td>
</tr>
<tr>
<td>( d )</td>
<td>OFDM symbols per RB</td>
<td>3</td>
</tr>
<tr>
<td>( L )</td>
<td>Number of used diversity branches</td>
<td>( \leq L_o )</td>
</tr>
<tr>
<td>( \ell_{\text{max}} )</td>
<td>Max. number of transmission rounds</td>
<td>([L_o/L])</td>
</tr>
</tbody>
</table>

### Table I

Parameters for TDL-C 300 ns–3 km/h (upper half) and signal (bottom half).

- **Symbol**: Represents various parameters associated with the channel model.
- **Parameter**: Describes the specific parameter being considered, such as bandwidth, system latency, and coherence time.
- **Value**: The numerical values assigned to each parameter, such as 0.66 MHz for the 50% coherence bandwidth.
practically-relevant pilot-based transmission schemes.

binary-input AWGN channel, to block-fading scenarios and
in [11], where the physical layer was modeled as a simple
introduced in this paper can be used to extend the analysis
fading channels with limited diversity. Specifically, the tools
sequence, the results in this paper are relevant to the joint coding-
of the latency rather than on its average [11]. As a conse-
An important application where the CDF of the delay plays
an optimum latency-aware design must be based on the CDF
remaining parameter values are given in Table I.
step function since the latency is deterministic whereas the CDF
of HARQ is a staircase function with steps at multiples of the
probability of the latency in HARQ being larger than in FBL-
Fig. 3. $E^*/N_0$ versus average latency for $L \in \{2, 3, 5, 6\}$ and 30 information
bits. The remaining parameter values are given in Table I.

Fig. 4. $R^*$ versus average latency for $\rho = -2$ dB and $L \in \{2, 3, 5, 6\}$. The
remaining parameter values are given in Table I.

Fig. 5. CDF of the latency. The curves are generated for 30 information bits,
$\rho = \{-5, -2\}$ dB and $L = 2$. The remaining parameter values are given in
Table I.

REFERENCES

low-latency wireless communication with short packets,” Proc. IEEE, vol. 104,
2359, May 2010.
antenna fading channels at finite blocklength,” IEEE Trans. Inf. Theory,
communications over multiple-antenna Rayleigh-fading channels,” IEEE
“Short packets over block-memoryless fading channels: Pilot-assisted
1712.06387.pdf
assisted short-packet transmission over multi-antenna fading channels: A
2011.
volutional coding for short blocklengths with decision feedback,” IEEE
[10] A. Avrantas, M. Kountouris, and P. Ciblat, “Energy-latency tradeoff in
ultra-reliable low-latency communication with retransmissions.”
Biyikoglu, “Reliable transmission of short packets through queues and
noisy channels under latency and peak-age violation guarantees.”
[12] A. Martinez and A. Guilian i Fàbregas, “Saddlepoint approximation of
random–coding bounds,” in Proc. Inf. Theory Applicat. Workshop (ITA),
1958.
[15] 3GPP, “TR 38.901: Study on channel model for frequencies from 0.5 to
[16] P. Schulz et al., “Latency critical IoT applications in 5G: Perspective on
the design of radio interface and network architecture,” IEEE Commun.