A New HARQ Scheme for 5G Systems via Interleaved Superposition Retransmission

Qianfan Wang[1,2], Li Chen[3,*], Xiao Ma[1]

[1] School of Computer Science and Engineering, Guangdong Key Laboratory of Information Security Technology, Sun Yat-sen University, Guangzhou 510006, China
[2] College of Computer Science, Jiaying University, Meizhou 514015, China
[3] School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou 510006, China
[*] The corresponding author, email: chenli55@mail.sysu.edu.cn

Cite as: Q. Wang, L. Chen, *et al*., "A new harq scheme for 5g systems via interleaved superposition retransmission," *China Communications*, vol. 20, no. 4, pp. 1-11, 2023. **DOI**: 10.23919/JCC.fa.2022-0670.202304

Abstract: Within the framework of the 5G new radio (NR), we propose a new hybrid automatic repeat request (HARQ) scheme to improve the throughput performance. The difference between the proposed scheme and the conventional one lies in the first retransmission, where the erroneous coded block group is interleaved and superimposed (XORed) onto a fresh coded block group. At the receiver, an iterative message-passing decoding algorithm can be employed to recover the target erroneous code block group (CBG). Only when the superposed retransmission fails, the conventional incremental redundancy (IR) or repetition redundancy (RR) retransmission is initiated. In any case, since the first retransmission is along with but has negligible effect on the fresh CBG, it costs neither transmitted power nor bandwidth. Monte-Carlo simulation results reveal that the presented HARQ schemes can achieve throughput improvements up to 10% over block fading channels and up to 50% over fast fading channels in comparison with the original 5G CBG-level HARQ scheme but without excessively increasing the implementation complexity.

Keywords: 5G NR; HARQ; LDPC codes; superposition retransmission

Received: Sep. 07, 2022
Revised: Nov. 11, 2022
Editor: Zhen Gao

I. INTRODUCTION

Hybrid automatic repeat request (HARQ) is a popularly used technique to ensure the reliable information transmission. In the conventional HARQ schemes, the receiver performs the decoding algorithm and then sends the black acknowledgement (ACK) or negative acknowledgement (NACK) feedback to the transmitter. Upon receiving an ACK, the transmitter starts a new transmission of the fresh data. While upon receiving a NACK, the transmitter would initiate a retransmission of the erroneous data, where the repetition redundancy (RR) [1] and the incremental redundancy (IR) [2] can be employed.

1.1 Background

In practical systems such as LTE and 5G new radio (NR), the transport block (TB)-based transmission is adopted, where a TB (medium access control (MAC) layer information) is segmented into multiple finite-length code blocks (CBs) and each CB with its cyclic redundancy check (CRC) bits is separately encoded into a coded block (codeword) using the channel coding at the physical layer. In particular, several CBs are grouped into a CB group (CBG), which is the retransmission unit for the 5G NR HARQ protocol. In the case when some CBs within a CBG are unsuccessfully recovered, the transmitter initiates retransmission of the entire CBG in an IR manner [3–5]. This CBG-level retransmission of the 5G standard trades off the resource efficiency of the transmission link and the overhead of the feedback link. However, it may lead to a waste of transmission power and bandwidth since one single unsuccessfully decoded CB will trigger retransmitting the entire CBG. Evidently, the CB-level HARQ schemes proposed in [6, 7] can avoid this waste, where only the erroneous CBs need to be retransmitted. However, this CB-level HARQ leads to an excessive amount of feedback overhead, especially for a TB with up to 152 CBs in NR, since the transmitter needs to know the feedback of each CB.

1.2 Related Works and Motivations

Throughput is a widely-used performance metric to evaluate the efficiency of a HARQ scheme in vast rich literatures, e.g., [8–11]. It is defined as the average number of successfully received information bits per transmitted symbol.

Generally, there are two different classes of approaches to improving the throughput. *i*) One is based on the single-packet design, in which typically fixed-length RR with Chase combining [12] and IR [2] can be employed for a single packet. Additionally, variable-length retransmission scheme for a single packet was considered in [8, 13, 14]. *ii*) The other is based on the multiple-packets design, where the redundancies can be generated from multiple data packets, e.g., the network-coded HARQ [15], superposition-coded HARQ [16], and compression-based HARQ [17].

The goal of this paper is to design a new HARQ scheme to improve the throughput performance based on the off-the-shelf 5G HARQ scheme. The original CBG-level retransmission needs to be maintained and the implementation complexity needs to be low for practical use. Given these requirements, we mainly consider the design of the encoder and decoder for the HARQ scheme. As a result, it is straightforward to integrate the presented HARQ scheme with 5G emerging enabling technologies, such as the multi-tier future cellular communication [18], the reconfigurable intelligent surface-assisted communication [19] and the Millimeter and THz communication [20, 21].

1.3 Basic Ideas and Contributions

Based on these above requirements, this paper proposed the superposition retransmission (SR) technique for the first retransmission of the erroneous CBG, resulting in the SR-HARQ scheme. This HARQ scheme is inspired by the idea of the block Markov superposition transmission [22, 23] and the work of [24–26]. Different from the partial superposition (PS)-HARQ scheme of [24], which is designed for the CB-level retransmission, the proposed SR-HARQ schemes are tailored to the CBG-level retransmission and they are more implementation friendly. Also distinguished from the PS-HARQ, the SR-HARQ circumvents optimizing the superposition fraction and employs the row-column interleaver, which is more hardware-friendly than the random construction of the PS-HARQ scheme. In the case when the first transmission of a CBG fails, the coded bits of this CBG are interleaved and superimposed onto a new coded block group, resulting in a mixed retransmission. At the receiver, an iterative message-passing algorithm is employed to recover the erroneous CBG. If the erroneous CBG is still unrecovered, the transceiver shifts into the original manner by sending IRs or RRs. Once the targeted CBG is recovered or estimated, its "interference" on the fresh CBG can be removed, resulting in a noisy reception for the fresh CBG at the receiver. In this sense, the mixed retransmission of the erroneous CBG costs neither extra bandwidth nor extra transmission power, implying a potential throughput gain. Based on the renewal-reward theorem, we derive the average throughput to analyze the throughput, confirming the throughput improvement over the conventional NR HARQ scheme. This throughput improvement is also confirmed by the Monte-Carlo simulations, showing that in comparison with the original 5G NR HARQ, throughput improvements can be up to 10% over block fading channels and 50% over fast fading channels.

II. PRELIMINARIES

2.1 System Model

Consider an end-to-end system, where we assume that data channel is noisy while the feedback channel is noiseless and delay-free. Taking binary phase-shift keying (BPSK) modulation as an example, we present the system model as follows. At slot $t \geqslant 0$, the transmitter sends a coded block group with length nB using BPSK modulation over a fading channel and the receiver obtains a noisy vector $\boldsymbol{y}^{(t)}$ expressed as

$$y_j^{(t)} = h_j^{(t)} x_j^{(t)} + w_j^{(t)}, \tag{1}$$

for $0 \leqslant j \leqslant nB-1$, where $x_j^{(t)} \in \{+1, -1\}$ is the j-th component of the transmitted vector $\boldsymbol{x}^{(t)}$, $y_j^{(t)}$ is the j-th component of the received vector $\boldsymbol{y}^{(t)}$, $w_j^{(t)}$ is a sample from an independent Gaussian random variable with distribution $N(0, \sigma^2)$, and $h_j^{(t)}$ is a sample from a Rayleigh distribution $\mathcal{R}$ with $\mathbb{E}[\mathcal{R}^2] = 1$. Assume that $h_j^{(t)}$, perfectly known to the receiver, keeps unchanged in a coherence period of T_f symbols and changes independently with an identical distribution at different coherence periods. Based on [27–29], we consider in this paper the following two set-ups: $T_f = 1$, referred to as fast fading channel, and $T_f = n/F$, referred to as block fading channel (where F denotes the number of fading blocks as [30]).

2.2 Review of NR HARQ Protocol

In the 5G NR system, a block of data from MAC layer during a single transmission time interval (TTI) is referred to as a TB. The TB with TB-CRC bits is segmented into M CBs and each CB is lengthened by appending its own CRC bits (referred to as CB-CRC bits). In particular, M CBs within a TB are evenly grouped into m CBGs so that each CBG contains at most $\lceil M/m \rceil$ and at least $\lfloor M/m \rfloor$ adjacent CBs. In this paper, for notational simplicity, we assume that each CBG consists of B CBs. After the CB segmentation, each CB with its corresponding CB-CRC bits is the input of the NR LDPC encoder for the data channel, resulting in the coded bits stored in a buffer for the transmission and the retransmission (if necessary). The NR LDPC codes are a kind of raptor-like rate-compatible LDPC codes [31]. The above procedure for segmenting a TB into multiple CBs and encoding CBs is illustrated in Figure 1.

For IR transmission manner, the CB with its CB-CRC bits is encoded by the lowest-rate code and the generating coded bits (except the first punctured $2Z$

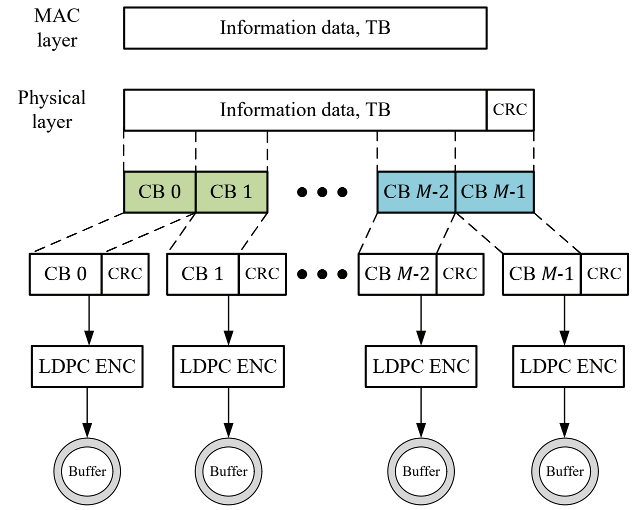

Figure 1. *The procedure for segmenting a TB into multiple CBs and encoding CBs, where the CBs with the same colour are grouped into a CBG.*

information bits) are written in a circular buffer. The coded bits in the buffer are divided into four (overlapping) transmission versions, denoted as $\boldsymbol{v}_{1\mathrm{T}}$, $\boldsymbol{v}_{2\mathrm{T}}$, $\boldsymbol{v}_{3\mathrm{T}}$ and $\boldsymbol{v}_{4\mathrm{T}}$, which are specified by a set of fixed locations, see Figure 2 for reference. In this circular buffer, $\boldsymbol{v}_{1\mathrm{T}}$ is transmitted in the first transmission round and the other three transmission versions are transmitted in sequence in the three retransmission rounds if necessary.

Given the noisy reception of a TB, the receiver decodes each CB and its CB-CRC bits. A CB is said to be unsuccessfully decoded if the decoded CB fails to reproduce the decoded CRC bits or the decoded vector is not a valid codeword. Once one unsuccessfully decoded CB is detected, the transmitter retransmits the whole CBG in an IR manner with NR LDPC codes, where the transmitter selects the corresponding retransmission versions (including $\boldsymbol{v}_{2\mathrm{T}}$, $\boldsymbol{v}_{3\mathrm{T}}$ and $\boldsymbol{v}_{4\mathrm{T}}$) according to the current transmission round.

III. THE SUPERPOSITION RETRANSMISSION HARQ

3.1 SR-RR-HARQ Scheme

Let $\mathscr{C}[n, k]$ be the LDPC code with length n and dimension k over the binary field $\mathbb{F}_2$. Let $\boldsymbol{u}^{(0)}, \boldsymbol{u}^{(1)}, \cdots, \boldsymbol{u}^{(L)}$ be a sequence of CBGs to be transmitted, where $\boldsymbol{u}^{(t)} = \left(\boldsymbol{u}^{(t,0)}, \boldsymbol{u}^{(t,1)}, \cdots, \boldsymbol{u}^{(t,B-1)}\right)$ and $\boldsymbol{u}^{(t,i)} \in \mathbb{F}_2^{k-k_c}$ for $0 \leqslant i \leqslant B-1$. At slot t, each $\boldsymbol{u}^{(t,i)}$ of $\boldsymbol{u}^{(t)}$

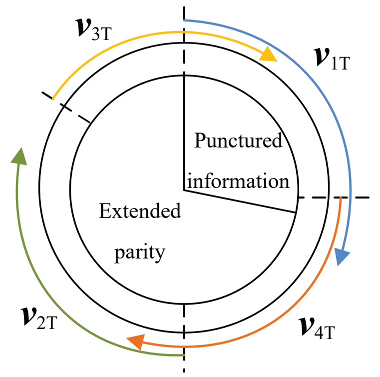

Figure 2. *An illustration of the 5G HARQ codeword circular buffer.*

with its CRC bits (of length k_c) is encoded into an LDPC codeword, resulting in a coded block group $\boldsymbol{v}^{(t)} = \left(\boldsymbol{v}^{(t,0)}, \boldsymbol{v}^{(t,1)}, \cdots, \boldsymbol{v}^{(t,B-1)}\right)$. In this letter, $\mathscr{C}[n,k]$ is considered as an NR LDPC code.

We take Figure 3 as an example to illustrate the difference between the conventional RR-HARQ and the presented SR-RR-HARQ, where we assume that $\boldsymbol{v}_{i\mathrm{T}}^{(t)} = \boldsymbol{v}^{(t)}$ for $i = 1, 2, \cdots$ for RR implementation. We focus on the transmission of $\boldsymbol{u}^{(0)}$ since the transmission of $\boldsymbol{u}^{(t)}$ with $t > 0$ is essentially the same. Assume that $\boldsymbol{u}^{(0)}$ is unrecovered (indicated by the CRC bits or the parity checks) after the first transmission and a NACK is received. In conventional RR-HARQ, see Figure 3(a) for reference, the transmitter resends $\boldsymbol{v}^{(0)}$ until an ACK is received or a maximum transmission number $T_{\max}$ is achieved. While in SR-RR-HARQ, see Figure 3(b) for reference, $\boldsymbol{v}^{(0)}$ is interleaved and superimposed onto $\boldsymbol{v}^{(1)}$, resulting in a mixed retransmission. To be precise, firstly, $\boldsymbol{v}^{(0)}$ are written in a row-column buffer with B rows and n columns row-by-row and read out column-by-column, resulting in the interleaved version of $\boldsymbol{v}^{(0)}$, denoted as $\widetilde{\boldsymbol{v}}^{(0)}$. For notational convenience, we use Π to denote the row-column interleaver and write down $\widetilde{\boldsymbol{v}}^{(0)} = \Pi(\boldsymbol{v}^{(0)})$. Secondly, the first retransmission for $\boldsymbol{v}^{(0)}$ can be generated as

$$\boldsymbol{c}^{(1)} = \widetilde{\boldsymbol{v}}^{(0)} + \boldsymbol{v}^{(1)}. \tag{2}$$

At the receiver, the noisy versions $\boldsymbol{y}^{(0)}$ of $\boldsymbol{v}^{(0)}$ and $\boldsymbol{y}^{(1)}$ of $\boldsymbol{c}^{(1)}$ can be utilized to recover $\boldsymbol{u}^{(0)}$ by an iterative message-passing decoding algorithm. The decoding algorithm is similar to that presented in [32] and is outlined in Algorithm 1, where $\boldsymbol{z}_{0\to1}$ and $\boldsymbol{z}_{1\to0}$, initialized

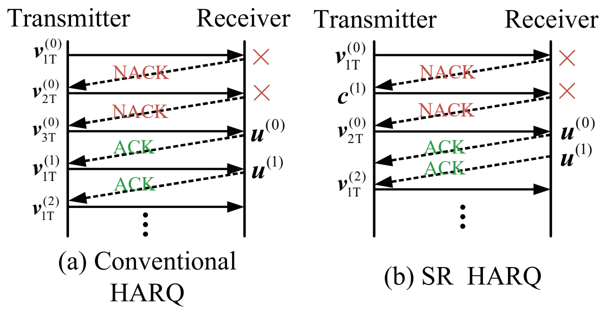

Figure 3. *An example of the conventional HARQ and the SR HARQ.*

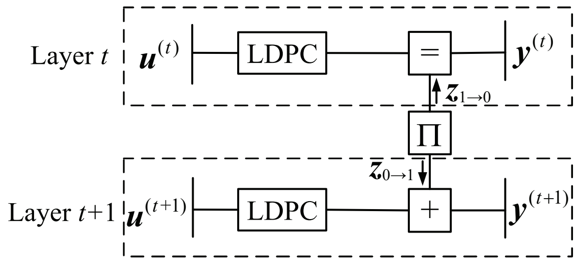

Figure 4. *Normal graph for the joint message-passing decoding, where the message passing at the node* [LDPC] *is implemented using a soft-in-soft-out (SISO) decoder with the maximum iteration number $I_{\max}$, the message passing at the node* [+] *is similar to that of the check node, the message passing at the node* [=] *is similar to that of the variable node, and the node* [Π]*, which is the row-column interleaver, simply transfers the messages between the node* [+] *and the node* [=].

by zero vectors, are messages exchanged between the two layers (see Figure 4 for reference). If $\boldsymbol{v}^{(0)}$ is still unrecovered, the transmitter switches into the original RR manner, where $\boldsymbol{v}^{(0)}$ is retransmitted until an ACK is received or $T_{\max}$ is achieved. It is worth pointing out that, in the case when the mixed retransmission $\boldsymbol{c}^{(1)}$ fails, the decoding will also switch into the original manner which does not involve the noisy version $\boldsymbol{y}^{(1)}$. This can further reduce the decoding complexity and has not been exploited for the PS-HARQ [24]. Once $\boldsymbol{v}^{(0)}$ is recovered or estimated, its "interference" on $\boldsymbol{y}^{(1)}$ can be removed, resulting in a noisy reception of $\boldsymbol{v}^{(1)}$. In this sense, transmitting $\boldsymbol{c}^{(1)}$ should be counted as a transmission for $\boldsymbol{v}^{(1)}$ rather than for $\boldsymbol{v}^{(0)}$. In other words, from the perspective of the transmission of $\boldsymbol{v}^{(0)}$, the mixed retransmission $\boldsymbol{c}^{(1)}$ helps to recover $\boldsymbol{v}^{(0)}$ but costs neither extra transmitted power nor bandwidth.

Considering the worst-case scenario and tak-

Algorithm 1. *Joint message-passing algorithm for decoding.*

Input: $\boldsymbol{y}^{(t)}, \boldsymbol{y}^{(t+1)}$ and the corresponding LLRs, also denoted as $\boldsymbol{y}^{(t)}, \boldsymbol{y}^{(t+1)}$

repeat

1) Taking into account the message-passing rules of the node $\boxed{\text{LDPC}}$ and the node $\boxed{=}$, the processor at layer t takes $\boldsymbol{y}^{(t)}$ and $\boldsymbol{z}_{1\to 0}$ as input, and delivers the extrinsic LLRs $\boldsymbol{z}_{0\to 1}$ to layer $t+1$;

2) Taking into account the message-passing rules of the node $\boxed{\text{LDPC}}$ and the node $\boxed{+}$, the processor at layer $t+1$ takes $\boldsymbol{y}^{(t+1)}$ and $\boldsymbol{z}_{0\to 1}$ as input, and delivers the extrinsic LLRs $\boldsymbol{z}_{1\to 0}$ to layer t;

3) Make hard decisions at layer t, resulting in $\hat{\boldsymbol{v}}^{(t)}$ and the corresponding $\hat{\boldsymbol{u}}^{(t)}$;

until $\hat{\boldsymbol{v}}^{(t)}$ is valid **or** a preset maximum global iteration number $J_{\max}$ is reached;

Output: $\hat{\boldsymbol{v}}^{(t)} = D(\boldsymbol{y}^{(t)}, \boldsymbol{y}^{(t+1)})$

Algorithm 2. *Transmission scheme of SR-RR-HARQ.*

Initialization: Set $t = 0$. The transmitter sends $\boldsymbol{v}^{(0)}$;

while $t \leqslant L$ **do**

  if $\hat{\boldsymbol{v}}^{(t)}$ *is successfully transmitted (indicated by an ACK of* $\boldsymbol{v}^{(t)}$*)* **then**

    The transmitter sends a fresh coded block group $\boldsymbol{v}^{(t+1)}$;

  else

    The transmitter sends a superposed coded block group $\boldsymbol{c}^{(t+1)} = \boldsymbol{v}^{(t+1)} + \Pi(\boldsymbol{v}^{(t)})$, where Π is the row-column interleaver;

    while $\hat{\boldsymbol{v}}^{(t)}$ *is unsuccessfully transmitted, and the transmission number of* $\boldsymbol{v}^{(t)}$ *is less than* $T_{\max}$ **do**

      The transmitter resends the repetition coded block group $\boldsymbol{v}^{(t)}$;

  $t \leftarrow t+1$;

ing $T_{\max} = 4$ as an example, we compare the SR-RR-HARQ scheme with the conventional RR-HARQ scheme by tabulating the transmitted coded block group in Table 1. In the worst case when the target packet is still failed after the $T_{\max}$ transmissions, it will be "threw away" as the standard 5G NR. The superposition is bit-wise XOR without lengthening the packet and the presented HARQ scheme can keep the transmission queue unchanged. For completely, the transmission scheme and reception scheme of the SR-RR-HARQ is summarized in Algorithm 2 and Algorithm 3, respectively.

3.2 SR-IR-HARQ Scheme

Additionally, the SR-HARQ scheme can be adapted to the IR mode, resulting in the SR-IR-HARQ scheme. Different from the SR-RR-HARQ scheme, the SR-IR-HARQ scheme retransmits IRs instead of RRs if necessary. For example, considering the worst case, the transmission for $\boldsymbol{u}^{(0)}$ is scheduled as $\{\boldsymbol{v}_{1\text{T}}^{(0)}, \boldsymbol{v}_{2\text{T}}^{(0)}, \boldsymbol{v}_{3\text{T}}^{(0)}, \boldsymbol{v}_{4\text{T}}^{(0)}\}$ for the 5G HARQ scheme and as $\{\boldsymbol{v}_{1\text{T}}^{(0)}, \boldsymbol{c}^{(1)}, \boldsymbol{v}_{2\text{T}}^{(0)}, \boldsymbol{v}_{3\text{T}}^{(0)}, \boldsymbol{v}_{4\text{T}}^{(0)}\}$ for the SR-IR-HARQ scheme, where $\boldsymbol{c}^{(1)} = \boldsymbol{v}_{1\text{T}}^{(1)} + \Pi(\boldsymbol{v}_{1\text{T}}^{(0)})$ is the same as that of the SR-RR-HARQ scheme. We also compare the SR-IR-HARQ scheme with the 5G IR-HARQ scheme, where the transmitted coded block groups for $(\boldsymbol{u}^{(0)}, \boldsymbol{u}^{(1)}, \cdots, \boldsymbol{u}^{(L)})$ in the worst case are also tabulated in Table 1. Also different from the SR-RR-HARQ, the SR-IR-HARQ recovers the unsuccessful CB with IRs (if any) over a larger Tanner graph associated with the lower rate LDPC code since the Chase combining is not directly applicable. The algorithm of the SR-IR-HARQ is omitted here.

3.3 Throughput Performance Analysis

As investigated in [33], the HARQ process can be modeled as a Markov process and the average throughput can be calculated by invoking the renewal-reward theorem, where the transmission rounds can be viewed as states and a HARQ cycle can be viewed as a renewal cycle. The average throughput is a ratio between the mean reward (the number of correctly decoded information bits per HARQ cycle) and the mean renewal time (the expected number of transmitted symbols per HARQ cycle). Let p_i denote the unsuccessful decoding probability in the i-th round after receiving the i-th redundancy versions of $\boldsymbol{u}^{(t)}$. Evidently, the probability of successful decoding in the i-th round is given by $p_{i-1} - p_i$. For the truncated NR HARQ scheme, where $T_{\max}$ is finite, the throughput can be calculated as

$$\tau_{\text{NR}} = \frac{R(1 - p_{T_{\max}})}{1 + p_1 + \sum_{i=2}^{T_{\max}-1} p_i}. \tag{3}$$

We need to point out that p_i with $i \geqslant 2$ of the NR HARQ scheme are similar to these of the presented HARQ due to the identical retransmission redundan-

Table 1. *Comparison of the conventional HARQ and SR-HARQ scheme in the worst-case scenario.*

	RR-HARQ	SR-RR-HARQ	(5G) IR-HARQ	SR-IR-HARQ
$t=0$	$\{\boldsymbol{v}^{(0)},\boldsymbol{v}^{(0)},\boldsymbol{v}^{(0)},\boldsymbol{v}^{(0)}\}$	$\{\boldsymbol{v}^{(0)},\boldsymbol{c}^{(1)},\boldsymbol{v}^{(0)},\boldsymbol{v}^{(0)},\boldsymbol{v}^{(0)}\}$	$\{\boldsymbol{v}_{1T}^{(0)},\boldsymbol{v}_{2T}^{(0)},\boldsymbol{v}_{3T}^{(0)},\boldsymbol{v}_{4T}^{(0)}\}$	$\{\boldsymbol{v}_{1T}^{(0)},\boldsymbol{c}^{(1)},\boldsymbol{v}_{2T}^{(0)},\boldsymbol{v}_{3T}^{(0)},\boldsymbol{v}_{4T}^{(0)}\}$
$t=1$	$\{\boldsymbol{v}^{(1)},\boldsymbol{v}^{(1)},\boldsymbol{v}^{(1)},\boldsymbol{v}^{(1)}\}$	$\{\boldsymbol{c}^{(2)},\boldsymbol{v}^{(1)},\boldsymbol{v}^{(1)},\boldsymbol{v}^{(1)}\}$	$\{\boldsymbol{v}_{1T}^{(1)},\boldsymbol{v}_{2T}^{(1)},\boldsymbol{v}_{3T}^{(1)},\boldsymbol{v}_{4T}^{(1)}\}$	$\{\boldsymbol{c}^{(2)},\boldsymbol{v}_{2T}^{(1)},\boldsymbol{v}_{3T}^{(1)},\boldsymbol{v}_{4T}^{(1)}\}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$t=L-1$	$\{\boldsymbol{v}^{(L-1)},\boldsymbol{v}^{(L-1)},\boldsymbol{v}^{(L-1)},\boldsymbol{v}^{(L-1)}\}$	$\{\boldsymbol{c}^{(L)},\boldsymbol{v}^{(L-1)},\boldsymbol{v}^{(L-1)},\boldsymbol{v}^{(L-1)}\}$	$\{\boldsymbol{v}_{1T}^{(L-1)},\boldsymbol{v}_{2T}^{(L-1)},\boldsymbol{v}_{3T}^{(L-1)},\boldsymbol{v}_{4T}^{(L-1)}\}$	$\{\boldsymbol{c}^{(L)},\boldsymbol{v}_{2T}^{(L-1)},\boldsymbol{v}_{3T}^{(L-1)},\boldsymbol{v}_{4T}^{(L-1)}\}$
$t=L$	$\{\boldsymbol{v}^{(L)},\boldsymbol{v}^{(L)},\boldsymbol{v}^{(L)},\boldsymbol{v}^{(L)}\}$	$\{\boldsymbol{v}^{(L)},\boldsymbol{v}^{(L)},\boldsymbol{v}^{(L)}\}$	$\{\boldsymbol{v}_{1T}^{(L)},\boldsymbol{v}_{2T}^{(L)},\boldsymbol{v}_{3T}^{(L)},\boldsymbol{v}_{4T}^{(L)}\}$	$\{\boldsymbol{v}_{2T}^{(L)},\boldsymbol{v}_{3T}^{(L)},\boldsymbol{v}_{4T}^{(L)}\}$

Algorithm 3. *Receiving scheme of SR-RR-HARQ at time* $t < L$.

Original Decoding: The receiver performs the decoding algorithm of $\mathscr{C}$ to obtain the estimation $\hat{\boldsymbol{v}}^{(t)}$ (also $\hat{\boldsymbol{u}}^{(t)}$) based on $\boldsymbol{y}^{(t)}$;

if $\hat{\boldsymbol{v}}^{(t)}$ *satisfies all parity checks and all CRC checks* **then**

- The receiver sends an ACK to the transmitter;

else

- The receiver sends a NACK to the transmitter;
- After receiving the noisy version $\boldsymbol{y}^{(t+1)}$ of $\boldsymbol{c}^{(t+1)}$, the receiver performs the joint message-passing decoding (Algorithm 1) with output $\hat{\boldsymbol{v}}^{(t)} = D(\boldsymbol{y}^{(t)}, \boldsymbol{y}^{(t+1)})$;
- **while** $\hat{\boldsymbol{v}}^{(t)}$ *is invalid indicated by the CB-CRC or the parity checks, and the transmission number of* $\boldsymbol{v}^{(t)}$ *is less than* $T_{\max}$ **do**
  - The receiver sends a NACK to the transmitter;
  - After receiving the noisy version $\boldsymbol{r}^{(t)}$ of $\boldsymbol{v}^{(t)}$, the receiver updates $\boldsymbol{y}^{(t)}$ as $\boldsymbol{y}^{(t)} \leftarrow \boldsymbol{y}^{(t)} + \boldsymbol{r}^{(t)}$ by Chase combining, and then performs the decoding algorithm of $\mathscr{C}$ with the updated $\boldsymbol{y}^{(t)}$.
- The receiver sends an ACK or a NACK to the transmitter and removes the effect of $\hat{\boldsymbol{v}}^{(t)}$ on $\boldsymbol{y}^{(t+1)}$. To be precise, $\boldsymbol{y}^{(t+1)} \leftarrow \boldsymbol{y}^{(t+1)} \odot (-1)^{\Pi(\hat{\boldsymbol{v}}^{(t)})}$, where the notation $\odot$ represents the component-wise multiplication;

cies. Similar to [25], the throughput of the SR-HARQ scheme can be estimated by

$$\tau_{\text{SR}} \approx \frac{R(1-p_{T_{\max}})}{1+\widetilde{p}_1+\sum_{i=2}^{T_{\max}-1} p_i}, \quad (4)$$

where $\widetilde{p}_1$ denotes the unsuccessful decoding probability after receiving the noisy versions of $\boldsymbol{v}^{(t)}$ and $\boldsymbol{c}^{(t+1)}$. In conventional BMST system [23] and BMST-LDPC codes [25], the error probability after receiving the noisy versions of $\boldsymbol{v}^{(t)}$ and $\boldsymbol{c}^{(t+1)}$ can be lower than that after receiving only the noisy version of $\boldsymbol{v}^{(t)}$. Therefore, based on equations (3)-(4), it is expected to achieve an enhanced throughput performance by using the presented HARQ scheme for $\widetilde{p}_1 < p_1$.

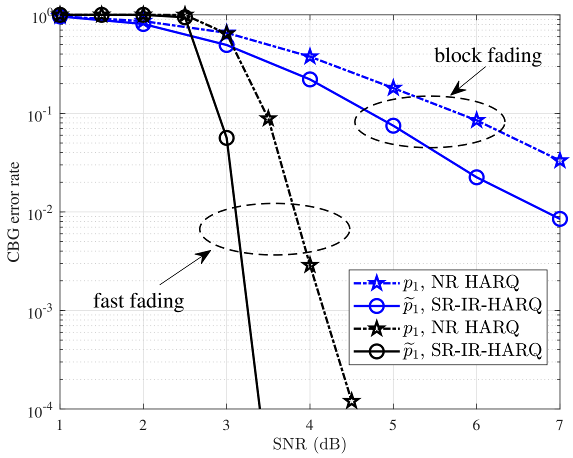

Figure 5. *CBG error rate performance over block fading channels and fast fading channels. A CBG has* $B = 16$ *(for block fading channels) or* $B = 8$ *CBs, where each CB of length* 696 *with its CB-CRC bits of length* 24 *is encoded by an NR LDPC code. For every transmission, the coded block length is* $n = 1440$.

To illustrate the error rate comparison, we consider BPSK modulation and the block fading channel with a coherence period of $T_f = 1440$. Each CB of length 696 along with its CB-CRC bits (of length 24) is then encoded by an NR LDPC code with dimension $k = 720$ (constructed from BG2 [4]). The results are shown in Figure 5, where we observe that $\widetilde{p}_1$ of the presented HARQ scheme can significantly less than p_1 of the 5G NR HARQ scheme.

IV. SIMULATION RESULTS

4.1 Throughput Comparison

In this subsection, we present the performance comparison between the SR-HARQ schemes and the 5G NR HARQ scheme over the independent Rayleigh fading channel using the Monte-Carlo simulations. The SR-RR-HARQ scheme is based on Algorithm 2 and Algorithm 3, and the SR-IR-HARQ scheme is based on the simple generalization of these Algorithms. Unless otherwise stated, we set $L = 100$. A CBG is consisting of B CBs, each of which is appended by 24 CRC bits generated with $g(x) = x^{24} + x^{23} + x^6 + x^5 + x + 1$ [4]. Each CB of length 696 along with its CB-CRC bits is then encoded by an NR LDPC code with dimension $k = 720$ (constructed from BG2 [4]), resulting in a codeword of length 3600 which is written in a circular buffer. All the transmission and retransmission for the 5G NR HARQ and SR-HARQ schemes have the length of $nB = 1440B$. For the presented SR-HARQ, we set $J_{\max} = 2$ and $I_{\max} = 25$ in the iterative message-passing decoding (Algorithm 1), where $J_{\max}$ denotes the global maximum iteration number and $I_{\max}$ denotes the local maximum iteration number of the SISO decoder implemented by the sum-product algorithm. For the NR HARQ, we set the maximum iteration number as 100 in the original LDPC decoding. The average throughput can be estimated by Monte-Carlo simulations as

$$\tau = \lim_{t\to\infty} \frac{K(t)}{N(t)}, \tag{5}$$

where $K(t)$ denotes the number of recovered information bits and $N(t)$ denotes the number of transmitted symbols, up to slot t. Evidently, with the aforementioned set-ups, the upper bound of the throughput is $(720 - 24)/1440 \approx 0.483$.

Example 1. *(Block fading channels): Consider the block fading channel and BPSK modulation, where the coherence period is $T_f = 1440$. Before modulation, for both the SR-HARQ scheme and the 5G NR HARQ scheme, we introduce a row-column interleaver of order nB for the B coded blocks. The comparison of the throughput performance is shown Figure 6, where we see that a throughput gain of up to 10% can be achieved by using the presented HARQ schemes over the original 5G IR HARQ scheme in the moderate-to-high SNR region. In addition, we see that the presented SR-HARQ schemes can outperform the PS-HARQ schemes presented in [24]. We also see that the proposed SR-RR-HARQ can perform as well as the SR-IR-HARQ in the moderate-to-high SNR region. While in the low SNR region, SR-RR-HARQ performs worse than the SR-IR-HARQ and the original 5G HARQ.*

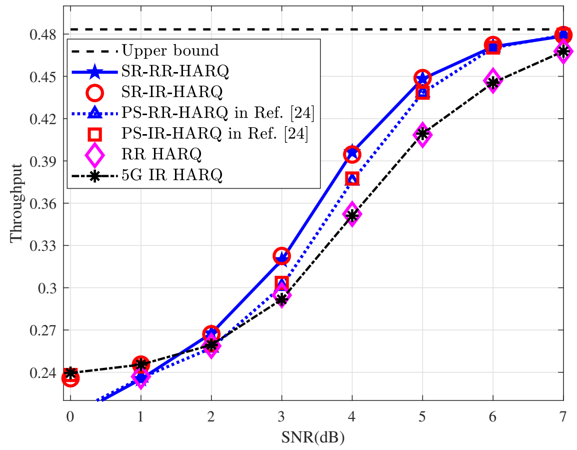

Figure 6. *Throughput performance over block fading channels. A CBG has $B = 16$ CBs and each CB of length 696 with its CB-CRC bits of length 24 is encoded by an NR LDPC code. For every transmission, the coded block length is $n = 1440$ and the coded block group length is $nB = 23040$. For PS-HARQ scheme, it is generalized to the CBG-level retransmission by considering a CBG with B CBs and the superposition fraction is $\alpha = 0.5$ as [24].*

We also present the throughput performance comparison with different fading blocks F as [30], where we set $L = 10$ and $B = 10$. The results are shown in Figure 7, from which we have similar observations. Additionally, we observe that the presented SR-IR-HARQ can outperform the original 5G IR HARQ, with different fading blocks F.

Example 2. *(Fast fading channels): Consider the fast fading channel and BPSK modulation, where the coherence period is $T_f = 1$. The comparison of the throughput performance over this fast fading channel is shown Figure 8, where we see that a throughput gain of up to 50% can be achieved by using the proposed HARQ schemes over the original 5G IR HARQ scheme in the moderate SNR region. We also see that the pre-*

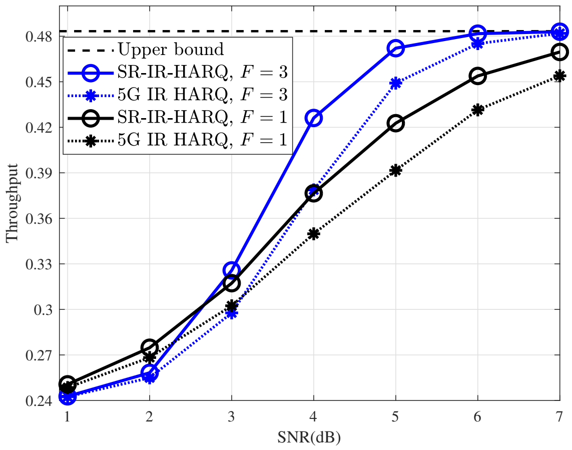

Figure 7. *Throughput performance over block fading channels with different fading blocks F for each coded block of length $n = 1440$. A CBG has $B = 10$ CBs and each CB of length 696 with its CB-CRC bits of length 24 is encoded by an NR LDPC code.*

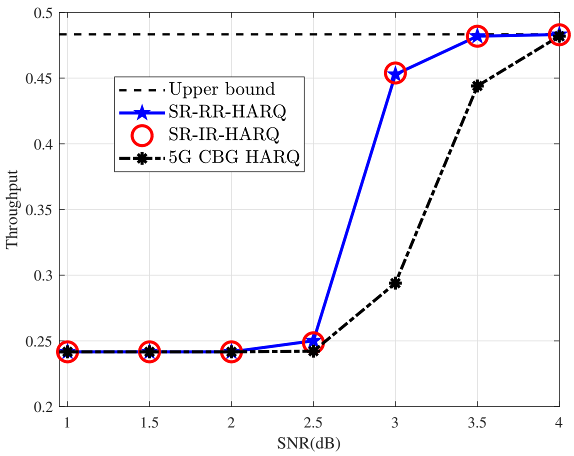

Figure 8. *Throughput performance over fast fading channels. A CBG has $B = 8$ CBs and each CB of length 696 with its CB-CRC bits of length 24 is encoded by an NR LDPC code. For every transmission, the coded block length is $n = 1440$ and the coded block group length is $nB = 11520$.*

sented SR-RR-HARQ can yield a similar performance to the SR-IR-HARQ.

Remark 1. *From the above simulation results, we observe that the presented SR-HARQ schemes can achieve greater throughput gain over fast fading channels than that over block fading channels. This is because the superposition retransmission is insufficient for the deep fading scenario, which occurs more likely in (non-ergodic) block fading channels than in (ergodic) fast fading channels.*

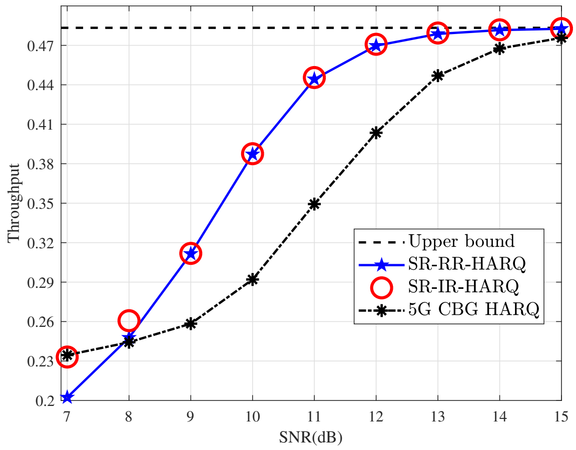

Figure 9. *Throughput performance over block fading channels with 16-QAM (Gray mapping) modulation. A CBG has $B = 16$ CBs and each CB of length 696 with its CB-CRC bits of length 24 is encoded by an NR LDPC code. For every transmission, the coded block length is $n = 1440$ and the coded block group length is $nB = 23040$.*

Example 3. *(16-QAM modulation): Consider the block fading channel and the 16-ary quadrature amplitude modulation (16-QAM), where the coherence period is $T_f = 360$. Before modulation, for both the SR-HARQ scheme and the 5G NR HARQ scheme, we introduce a row-column interleaver of order nB for the B coded blocks. The comparison of the throughput performance is shown Figure 9, where we see that a throughput gain of up to 10% can be achieved by using the presented HARQ schemes over the original 5G IR HARQ scheme in the high SNR region. We also see that the proposed SR-RR-HARQ can perform as well as the SR-IR-HARQ in the high SNR region. While in the low SNR region, SR-RR-HARQ performs worse than the SR-IR-HARQ and the original 5G HARQ.*

4.2 Complexity Analysis

The increased complexity of the presented schemes in comparison with the 5G NR HARQ scheme can be analyzed by taking as an example the transmission of $\boldsymbol{v}^{(0)}$. If $\boldsymbol{v}^{(0)}$ is recovered for the first transmission, the transceiver does not require extra computational overhead. Otherwise, the transmitter requires extra nB XORs and the receiver requires to perform a message-passing decoding algorithm with an itera-

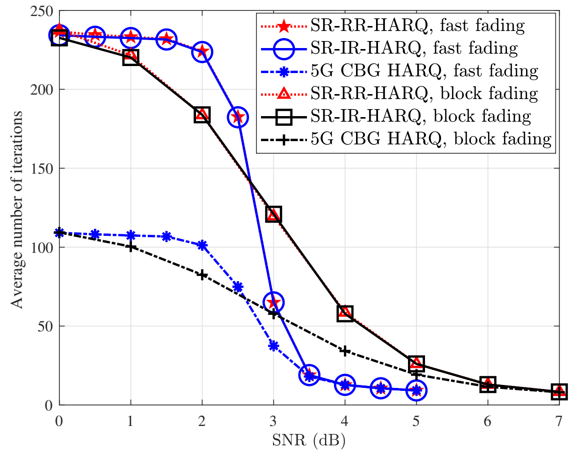

Figure 10. *The average iteration numbers for performing the LDPC decoding for each CB. The numerical results correspond to example 1 and example 2. For the SR-HARQ schemes, we set* $J_{\max} = 2$ *and* $I_{\max} = 25$*. For the 5G NR HARQ scheme, we set the maximum iteration number as 100.*

tive manner. We can measure the decoding complexity in terms of the average iteration numbers for decoding each CB. In particular, the parity-check and CRC-aided stopping criterion is employed in both the SR-HARQ and the 5G NR HARQ schemes. The comparison of the average iteration numbers is shown in Figure 10, where we see that the SR-IR-HARQ has a slightly less number of iterations as compared with the SR-RR-HARQ in the low SNR region. Furthermore, we see that the SR-HARQ incurs a large iteration number increase in the low SNR region, while they have a similar iteration number in the high SNR region, in comparison with the 5G NR HARQ scheme[1].

V. CONCLUSION

This letter has presented the throughput-enhanced CBG HARQ schemes based on the original 5G HARQ scheme, where a tailored first retransmission using the SR technique is introduced. The presented HARQ schemes can be easily implemented from the basic components of the existing 5G HARQ scheme. Monte-Carlo simulation results have revealed that, both the SR-IR-HARQ scheme and the simple SR-RR-HARQ scheme can achieve throughput improvements up to 10% over block fading channels and 50% over fast fading channels, in comparison with the 5G HARQ scheme.

ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China (No. 61971454 and No. 62071498) and the Guangdong Basic and Applied Basic Research Foundation (No. 2023A1515011056).

NOTES

[1]Due to this fact, the presented HARQ scheme may not be suitable in the low SNR region, especially for the delay sensitive applications.

References

[1] G. Benelli, "An ARQ scheme with memory and soft error detectors," *IEEE Transactions on Communications*, vol. 33, no. 3, pp. 285–288, Mar. 1985.

[2] D. M. Mandelbaum, "An adaptive-feedback coding scheme using incremental redundancy," *IEEE Transactions on Information Theory*, vol. 20, no. 3, pp. 388–389, May 1974.

[3] J. Yeo, T. Kim, *et al.*, "Advanced data transmission framework for 5G wireless communications in the 3GPP new radio standard," *IEEE Communications Standards Magazine*, vol. 3, no. 3, pp. 38–43, Sept. 2019.

[4] 3GPP, Multiplexing and channel coding (Release 15), 3GPP TS 38.212, 2018.

[5] 3GPP, Physical layer procedures for control (Release 15), 3GPP TS 38.213, 2020.

[6] H. Pai, Y. Han, *et al.*, "New HARQ scheme based on decoding of tail-biting convolutional codes in IEEE 802.16e," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 3, pp. 912–918, Mar. 2011.

[7] K. Pedersen, G. Pocovi, *et al.*, "Preemptive scheduling of latency critical traffic and its impact on mobile broadband performance," in *IEEE 87th Vehicular Technology Conference*, pp. 1–6, Porto, Portugal, 2018.

[8] L. Szczecinski, S. R. Khosravirad, *et al.*, "Rate allocation and adaptation for incremental redundancy truncated HARQ," *IEEE Transactions on Communications*, vol. 61, no. 6, pp. 2580–2590, Jun. 2013.

[9] P. Larsson, L. K. Rasmussen, *et al.*, "Throughput analysis of ARQ schemes in Gaussian block fading channels," *IEEE Transactions on Communications*, vol. 62, no. 7, pp. 2569–2588, Jul. 2014.

[10] H. Khoshnevis, I. Marsland, *et al.*, "Throughput-based design for polar-coded modulation," *IEEE Transactions on Communications*, vol. 67, no. 3, pp. 1770–1782, Mar. 2019.

[11] A. Ahmed, A. Al-Dweik, *et al.*, "Hybrid automatic repeat request (HARQ) in wireless communications systems and standards: A contemporary survey," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 4, pp. 2711–2752, 2021.

[12] 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.

[13] J. Cheng, Y. Wang, *et al.*, "Adaptive incremental redundancy," in *IEEE 58th Vehicular Technology Conference*, pp. 737–741, Orlando, FL, USA, Oct. 2003.
[14] G. Qiu, M. Zhao, *et al.*, "Throughput maximization for polar coded IR-HARQ using deep reinforcement learning," in *IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications*, pp. 1–6, London, UK, 2020.
[15] Y. Lang, D. Wubben, *et al.*, "Improved HARQ based on network coding and its application in LTE," in *IEEE Wireless Communications and Networking Conference*, pp. 1958–1963, Shanghai, China, Apr. 2012.
[16] 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.
[17] Y. Fu and H. Yang, "Compressed retransmission for cbg-harq," *IEEE Communications Letters*, vol. 26, no. 11, pp. 2522–2526, 2022.
[18] S. W. H. Shah, M. M. U. Rahman, *et al.*, "Effective capacity analysis of HARQ-enabled D2D communication in multi-tier cellular networks," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 9, pp. 9144–9159, 2021.
[19] S. W. H. Shah, A. N. Mian, *et al.*, "Statistical QoS analysis of reconfigurable intelligent surface-assisted D2D communication," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 7, pp. 7343–7358, 2022.
[20] T. S. Rappaport, Y. Xing, *et al.*, "Overview of millimeter wave communications for fifth-generation (5G) wireless networks-with a focus on propagation models," *IEEE Transactions on Antennas and Propagation*, vol. 65, no. 12, pp. 6213–6230, 2017.
[21] J. Zhang, K. Kang, *et al.*, "Millimeter and THz wave for 5G and beyond," *China Communications*, vol. 16, no. 2, pp. iii–vi, 2019.
[22] X. Ma, C. Liang, *et al.*, "Obtaining extra coding gain for short codes by block Markov superposition transmission," in *IEEE International Symposium on Information Theory*, pp. 2054–2058, Istanbul, Turkey, 2013.
[23] X. Ma, C. Liang, *et al.*, "Block Markov superposition transmission: Construction of big convolutional codes from short codes," *IEEE Transactions on Information Theory*, vol. 61, no. 6, pp. 3150–3163, Jun. 2015.
[24] Q. Wang, S. Cai, *et al.*, "A throughput-enhanced HARQ scheme for 5G system via partial superposition," *IEEE Communications Letters*, vol. 24, no. 10, pp. 2162–2166, Oct. 2020.
[25] Q. Wang, S. Cai, *et al.*, "Spatially coupled LDPC codes via partial superposition and their application to HARQ," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 4, pp. 3493–3504, Apr. 2021.
[26] Q. Wang, S. Cai, *et al.*, "Free-ride feedback and superposition retransmission over LDPC coded links," *IEEE Transactions on Communications*, vol. 71, no. 1, pp. 13–25, Jan. 2023.
[27] S. Muller-Weinfurtner, "Coding approaches for multiple antenna transmission in fast fading and OFDM," *IEEE Transactions on Signal Processing*, vol. 50, no. 10, pp. 2442–2450, Oct. 2002.
[28] J. J. Boutros, A. Guillen i Fabregas, *et al.*, "Low-density parity-check codes for nonergodic block-fading channels," *IEEE Transactions on Information Theory*, vol. 56, no. 9, pp. 4286–4300, 2010.
[29] B. Makki and T. Eriksson, "On the performance of MIMO-ARQ systems with channel state information at the receiver," *IEEE Transactions on Communications*, vol. 62, no. 5, pp. 1588–1603, May 2014.
[30] Y. Li and M. Salehi, "Quasi-cyclic LDPC code design for block-fading channels," in *44th Annual Conference on Information Sciences and Systems*, pp. 1–5, Princeton, NJ, USA, 2010.
[31] H. Li, B. Bai, *et al.*, "Algebra-assisted construction of quasi-cyclic LDPC codes for 5G new radio," *IEEE Access*, vol. 6, pp. 50 229–50 244, 2018.
[32] Q. Wang, S. Cai, *et al.*, "Spatially coupled LDPC codes via partial superposition," in *IEEE International Symposium on Information Theory*, pp. 2614–2618, Paris, France, Jul. 2019.
[33] 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, Jul. 2001.

Biographies

Qianfan Wang received the Ph.D. degrees from Sun Yat-sen University, China, in 2022. He is currently a Post-Doctoral Fellow with Sun Yat-sen University. His research interests include information theory, channel coding theory, and their applications to communication systems.

Li Chen received the B.Sc. degree in applied physics from Jinan University, China, in 2003, and the M.Sc. degree in communications and signal processing and the Ph.D. degree in communications engineering from Newcastle University, U.K., in 2004 and 2008, respectively. From 2007 to 2010, he was a Research Associate with Newcastle University. In 2010, he returned to China as a Lecturer of the School of Information Science and Technology, Sun Yat-sen University, Guangzhou. From 2011 to 2012, he was a Visiting Researcher with the Institute of Network Coding, the Chinese University of Hong Kong, China. From 2011 and 2016, he was an Associate Professor and a Professor of university. Since 2013, he has been the Associate Head of the Department of Electronic and Communication Engineering (ECE). From July 2015 to October 2015, he was a Visitor of the Institute of Communications Engineering, Ulm University, Germany. From October 2015 to June 2016, he was a Visiting Associate Professor with the Department of Electrical Engineering, University of Notre Dame, USA. From 2017 to 2020, he was the Deputy Dean of the School of Electronics and Communication Engineering. His research interests include information theory, error-correction codes, and data communications. He likes reading and photography. He is a Senior Member of the Chinese Institute of Electronics (CIE). He is a member of the IEEE Information Theory Society Board of Governors Conference Committee and External Nomination Committee, the Chair of the IEEE Information Theory Society Guangzhou Chapter, and a Committee Member of the CIE Information Theory Society. He is currently serving as an Associate Editor for IEEE Transactions on Communications. He has been involved in organizing several international conferences, including the 2018 IEEE Information Theory Workshop (ITW) at Guangzhou, for which he was the General Co-Chair.

Xiao Ma received the Ph.D. degree in communication and information systems from Xidian University, China, in 2000. He is a Professor of the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China. From 2000 to 2002, he was a Post-Doctoral Fellow with Harvard University, Cambridge, MA. From 2002 to 2004, he was a Research Fellow with City University of Hong Kong, China. His research interests include information theory, channel coding theory and their applications to communication systems and digital recording systems. He is a co-recipient, with A. Kavčić and N. Varnica, of the 2005 IEEE Best Paper Award in Signal Processing and Coding for Data Storage.