PAPER

# Highly-Efficient Low-Latency HARQ Built on NOMA for URLLC: Radio Resource Allocation and Transmission Rate Control Aspects*

**Ryota KOBAYASHI**[†], *Nonmember*, **Yasuaki YUDA**[††], *Member*, *and* **Kenichi HIGUCHI**[†a)], *Senior Member*

**SUMMARY** Hybrid automatic repeat request (HARQ) is an essential technology that efficiently reduces the transmission error rate. However, for ultra-reliable low latency communications (URLLC) in the 5th generation mobile communication systems and beyond, the increase in latency due to retransmission must be minimized in HARQ. In this paper, we propose a highly-efficient low-latency HARQ method built on non-orthogonal multiple access (NOMA) for URLLC while minimizing the performance loss for coexisting services (use cases) such as enhanced mobile broadband (eMBB). The proposed method can be seen as an extension of the conventional link-level non-orthogonal HARQ to the system-level protocol. This mitigates the problems of the conventional link-level non-orthogonal HARQ, which are decoding error under poor channel conditions and an increase in transmission delay due to restrictions in retransmission timing. In the proposed method, delay-sensitive URLLC packets are preferentially multiplexed with best-effort eMBB packets in the same channel using superposition coding to reduce the transmission latency of the URLLC packet while alleviating the throughput loss in eMBB. This is achieved using a weighted channel-aware resource allocator (scheduler). The inter-packet interference multiplexed in the same channel is removed using a successive interference canceller (SIC) at the receiver. Furthermore, the transmission rates for the initial transmission and retransmission are controlled in an appropriate manner for each service in order to deal with decoding errors caused by error in transmission rate control originating from a time varying channel. We show that the proposed method significantly improves the overall performance of a system that simultaneously provides eMBB and URLLC services.

***key words:*** *hybrid ARQ, non-orthogonal multiple access, NOMA, scheduling, transmission rate control, URLLC*

## 1. Introduction

In contrast to fourth-generation systems such as the Long Term Evolution (LTE) and LTE-Advanced [1], [2] where the primary service offered is mobile broadband, the fifth-generation New Radio (5G NR) system [3] and beyond [4] are expected to support a wider range of wireless communication services (use cases) such as massive machine-type communications (mMTC), ultra-reliable low latency communications (URLLC), and enhanced mobile broadband (eMBB). This paper focuses on the downlink cellular system in which eMBB and URLLC coexist and investigates a method for improving the low error rate and low transmission latency characteristics of URLLC while suppressing the deterioration in the system throughput for eMBB.

The hybrid automatic repeat request (HARQ) protocol [5] using error detection coding and powerful error correction coding such as the turbo code and low density parity check (LDPC) code efficiently achieves low error-rate transmission. Therefore, the effective use of HARQ is promising in efficiently achieving the low error rate of URLLC. However, increased transmission latency when packet retransmission is conducted must be addressed for URLLC. As an approach to address this problem, members of our research group reported on a low latency HARQ method that uses channel state information (CSI) prior to channel decoding [6]–[8]. This method mitigates the increased transmission latency by requesting early retransmission before the channel decoding process is completed based on the CSI obtained prior to channel decoding.

On the other hand, a non-orthogonal HARQ method in which the retransmission packet and the subsequent initial packet are non-orthogonally multiplexed in the same channel based on superposition coding is reported in [9]–[11]. This method reduces the bandwidth loss (throughput loss) associated with retransmission compared to that for the conventional orthogonal HARQ that allocates an exclusive channel (time-frequency slot) to the retransmission packet. Furthermore, assuming URLLC, non-orthogonal HARQ reduces the transmission latency since retransmission does not incur a transmission delay for the subsequent packet [10]–[12]. We note that applying non-orthogonal HARQ using superposition coding to the low latency HARQ method that uses CSI prior to channel decoding is investigated in [8].

However, non-orthogonal HARQ using superposition coding causes interference between superposition-coded packets. On the receiver side, this inter-packet interference is suppressed using a successive interference canceller (SIC) [10], [11]. Originally, HARQ is a link-level data transmission protocol, and so the HARQ protocol is defined between one pair of transmitter and receiver. This also holds true for the non-orthogonal HARQ reported in [9]–[12]. Therefore, the retransmission packet of some terminal is non-orthogonally multiplexed with the next packet for initial transmission of that terminal in the same channel. In this case however, when non-orthogonal HARQ is applied to a user terminal experiencing poor instantaneous channel conditions, the inter-packet interference canceling process us-

Manuscript received December 10, 2022.
Manuscript revised April 17, 2023.
Manuscript publicized May 22, 2023.

[†] The authors are with the Department of Electrical Engineering, Graduate School of Science and Technology, Tokyo University of Science, Noda-shi, 278-8510 Japan.

[††] The author is with Panasonic Holdings Corporation, Yokohama-shi, 224-8539 Japan.

[*] The material in this paper was presented in part at the 94th IEEE Vehicular Technology Conference (VTC2021-Fall), Virtual conference, 2021.

a) E-mail: higuchik@rs.tus.ac.jp

DOI: 10.1587/transcom.2022EBP3203

ing the SIC does not work well and the throughput may deteriorate. Non-orthogonal HARQ within a pair of transmitter and receiver, thus the link-level non-orthogonal HARQ, also limits the timing of retransmission. This makes it difficult to suppress the increase in transmission delay time associated with retransmissions.

To mitigate this problem and apply non-orthogonal HARQ to achieve a low error rate and low latency in URLLC in an efficient manner, we propose extending the link-level non-orthogonal HARQ in [9]–[12] to the system-level one. In cellular systems, a base station associates with multiple user terminals in a cell, and packets of different terminals are multiplexed based on scheduling within the framework of multiple access defined in the system. Therefore, when non-orthogonal HARQ using superposition coding is applied to a cellular system using non-orthogonal multiple access (NOMA) with a SIC [13], [14] as the multiple access scheme, retransmission packets can be non-orthogonally multiplexed with the packets of other terminals. The proposed non-orthogonal HARQ built on NOMA enables more effective use of non-orthogonal HARQ at the system level. As known from the theory of NOMA with a SIC [13], its effect can be greatly enhanced by non-orthogonal multiplexing of packets of user terminals under appropriately different channel states. Extending this idea, non-orthogonal multiplexing of retransmission packets and packets of other terminals in appropriately different channel states in terms of the SIC process within the same channel reduces the non-orthogonal HARQ problem at the link level and provides highly-efficient low-latency transmission for URLLC. In the proposed NOMA-based highly-efficient low-latency HARQ method, the radio resource allocation, including bandwidth and power allocation to selected terminals for non-orthogonal multiplexing, and the transmission rate control are key issues. This paper is a particularly detailed study of this point. There are many reports investigating system-level resource allocation and the HARQ method for URLLC, e.g., in [15]–[18]. However, these reports assume orthogonal multiple access (OMA). The proposed method is based on NOMA for the purpose of effectively utilizing non-orthogonal HARQ at the system level so that the requirements for an extremely low error rate and short delay time in URLLC are satisfied while the negative impact on the eMBB performance is alleviated.

The proposed method uses channel-aware time/frequency-domain instantaneous resource allocation including scheduling that multiplexes the URLLC packet to the best-effort eMBB packet within the same time-frequency block using superposition coding. In scheduling, it is important to obtain multi-user diversity between terminals by taking into account the instantaneous channel conditions of all candidate user terminals in all services. To achieve this, the proposed resource allocation metric is a throughput-based one for multiple services, which is based on [19], with service-dependent appropriate weighting, which is based on [19], [20], so that a low error rate and low transmission latency in URLLC can be explicitly guaranteed. Furthermore, in order to suppress the decoding error at the receiver caused by the error in transmission rate control due to the outdated CSI, i.e., channel quality indicator (CQI), we propose introducing service-dependent backoff operation in the transmission rate control. With the above configuration, the error rate and transmission delay time of URLLC are improved while suppressing the deterioration in the system throughput for eMBB operated at the same time. Extensive computer simulation results quantitatively show the performance gain when using the proposed method. We note that the proposed method can be operated in conjunction with the low-latency HARQ method in [6]–[8] using CSI before channel decoding. However, this paper assumes normal HARQ in which the retransmission request is sent after channel decoding is completed to assess the basic performance of the proposed system. We also note that the contents of this paper are based on [21], but include enhanced evaluation and discussions.

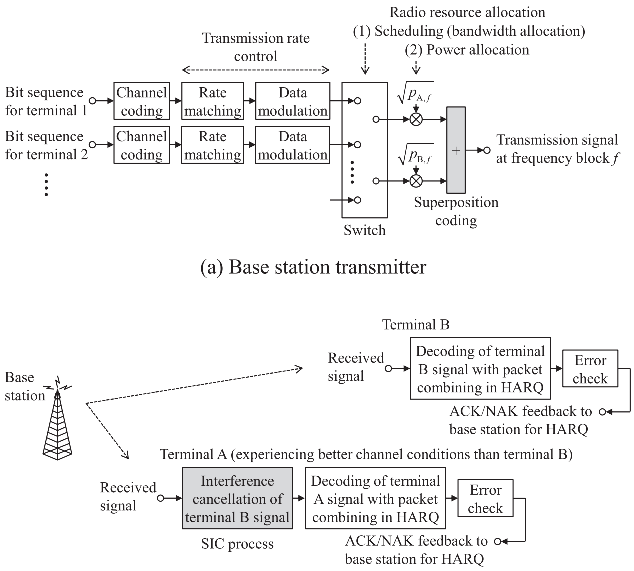
The remainder of this paper is organized as follows. First, Sect. 2 presents the proposed method. Section 3 shows the numerical results based on computer simulations. Finally, Sect. 4 concludes the paper.

## 2. Proposed Method

Since the proposed method is applied independently to each base station (cell), the control in the cell of interest is described below. It is assumed that eMBB and URLLC are operated simultaneously in the downlink in the cell of interest. The set of terminals that receive eMBB service in the cell is denoted as $\mathcal{K}_{\mathrm{eMBB}}$. The set of terminals that receive URLLC service in the cell is denoted as $\mathcal{K}_{\mathrm{URLLC}}$. The set of all terminals in the cell is denoted as $\mathcal{K} = \mathcal{K}_{\mathrm{eMBB}} \cup \mathcal{K}_{\mathrm{URLLC}}$. The set of downlink frequency blocks that is shared by eMBB and URLLC is denoted as $\mathcal{F}$.

In order to achieve ultra-high quality and a short delay at the same time in URLLC, the URLLC terminals must be prioritized in terms of the resources (bandwidth and power) allocated to the initial and retransmission packets. In addition, in order to suppress packet decoding errors at the user terminal receiver due to channel fluctuations over time, it is necessary to reduce the transmission rate to a value less than the maximum decodable rate suggested by the CQI report when determining the transmission rate. This leads to an increase in the bandwidth allocation per packet. When OMA allocates the time-frequency block orthogonally to each terminal, the bandwidth allocated to the best-effort-type eMBB is significantly limited in order to guarantee the QoS of URLLC, which causes severe throughput deterioration for eMBB.

To address this problem, in the proposed method, URLLC-prioritized resource allocation (including scheduling) is performed in conjunction with NOMA with a SIC, and eMBB packets can be multiplexed within the same time-frequency block using superposition coding to achieve a better tradeoff between ultra-high quality and a short delay in URLLC and a high throughput in eMBB. Figure 1 shows the


Radio resource allocation
(1) Scheduling (bandwidth allocation)
(2) Power allocation
Transmission rate control
Bit sequence for terminal 1
Channel coding
Rate matching
Data modulation
Bit sequence for terminal 2
Channel coding
Rate matching
Data modulation
$\sqrt{p_{A,f}}$
$\sqrt{p_{B,f}}$
+
Transmission signal at frequency block $f$
Superposition coding
Switch
Terminal B
Received signal
Decoding of terminal B signal with packet combining in HARQ
Error check
ACK/NAK feedback to base station for HARQ
Base station
Terminal A (experiencing better channel conditions than terminal B)
Received signal
Interference cancellation of terminal B signal
Decoding of terminal A signal with packet combining in HARQ
Error check
SIC process
ACK/NAK feedback to base station for HARQ

(a) Base station transmitter

(b) User terminal receiver

**Fig. 1** Base station transmitter and user terminal receiver in proposed method.

base station transmitter using superposition coding and the user terminal receiver using the SIC with HARQ in the proposed method. We note that Fig. 1 shows the process at one frequency block of interest. However, since the coded bit sequence of each user terminal is transmitted using multiple frequency blocks assigned to that terminal, the decoding process at the terminal is conducted jointly among multiple frequency blocks.

The proposed resource allocation metric is based on the throughput-based metric for multiple services reported in [19]. More specifically, assuming proportional fairness among eMBB user terminals, among URLLC user terminals, and between the two services, the resource allocation metric, $\lambda_f(\mathcal{U}_f, \mathcal{P}_f; t)$, to be maximized for frequency block $f \in \mathcal{F}$ at time $t$ is represented as

$$\lambda_f\left(\mathcal{U}_f, \mathcal{P}_f; t\right) = \sum_{k \in \mathcal{K}} \frac{r_{k,f}^{\mathrm{TX}}\left(\mathcal{U}_f, \mathcal{P}_f; t\right)}{\left|\mathcal{K}_{i_k}\right| R_k(t-1)}. \tag{1}$$

Here, $R_k(t)$ is the average throughput of terminal $k$ at time $t$. Term $r_{k,f}^{\mathrm{TX}}(\mathcal{U}_f, \mathcal{P}_f; t)$ is the transmission rate to terminal $k$ at frequency $f$, which corresponds to the expectation of the instantaneous throughput (channel capacity) assuming NOMA with a SIC for a given set of scheduled terminals, $\mathcal{U}_f$, and that of allocated power levels, $\mathcal{P}_f$, for terminals in $\mathcal{U}_f$. Term $i_k$ is the service indicator of user terminal $k$; thus, $i_k$ = 'eMBB' when $k \in \mathcal{K}_{\mathrm{eMBB}}$, while $i_k$ = 'URLLC' when $k \in \mathcal{K}_{\mathrm{URLLC}}$.

However, if the metric in (1) is used without alteration, the frequency block allocation to the initial or retransmission packet of the URLLC terminal cannot be given sufficient priority. Therefore, in order to maintain the multi-user diversity effect between terminals obtained by the metric in (1) as much as possible, and to guarantee explicitly a low error rate and short transmission delay time in URLLC at the same time, the weighting approach discussed, e.g., in [20] is applied between the services of the metric in (1). Specifically, non-negative constants $\delta_{\mathrm{eMBB}}$ and $\delta_{\mathrm{URLLC}}$ are defined as weights to be multiplied by the metrics of the eMBB terminal and URLLC terminal, respectively. The proposed resource allocation metric is represented as

$$\lambda_f\left(\mathcal{U}_f, \mathcal{P}_f; t\right) = \sum_{k \in \mathcal{K}} \frac{\delta_{i_k} r_{k,f}^{\mathrm{TX}}\left(\mathcal{U}_f, \mathcal{P}_f; t\right)}{\left|\mathcal{K}_{i_k}\right| \mathcal{R}_k(t-1)}. \tag{2}$$

By setting $\delta_{\mathrm{URLLC}}$ greater than $\delta_{\mathrm{eMBB}}$, the allocation of radio resources to URLLC terminals is prioritized. However, since the transmission rate, $r_{k,f}^{\mathrm{TX}}(\mathcal{U}_f, \mathcal{P}_f; t)$, depends on the instantaneous channel, the metric in (2) that takes this into consideration can obtain multi-user diversity gain across services, and in particular contributes to an increase in the system throughput for eMBB. In this paper, power distribution $\mathcal{P}_f$ between non-orthogonally multiplexed terminals is determined based on the fixed power allocation method in [22], and frequency block $f$ is assigned to terminal set $\mathcal{U}_f$ that maximizes the metric in (2).

Let $g_{k,f}(t)$ be the instantaneous channel gain normalized by the noise power including the inter-cell interference power at frequency block $f$ of terminal $k$ at time $t$. The set of terminals to which frequency block $f$ is assigned by the scheduler is denoted as $\mathcal{U}_f(t) \subseteq \mathcal{K}$. Here, $|\mathcal{U}_f(t)| = 1$ in OMA and $|\mathcal{U}_f(t)| \geq 1$ in NOMA. The transmission power assigned to terminal $k \in \mathcal{U}_f(t)$ is denoted as $p_{k,f}(t) \in \mathcal{P}_f(t)$.

A SIC is implemented in terminal receiver $k$ to remove the inter-terminal interference from the signal to terminal $j \neq k$, $j \in \mathcal{U}_f(t)$, depending on $g_{k,f}(t)$ and $g_{j,f}(t)$. The terminal order of decoding in the SIC is in the order of the normalized channel gain, $g_{k,f}(t)$ [13], [23]. Thus, terminal $k$ removes the inter-terminal interference from terminal $j$ whose $g_{j,f}(t)$ is lower than $g_{k,f}(t)$. The maximum transmission rate that terminal $k$ can correctly decode at frequency block $f$ at time $t$ for given $\mathcal{U}_f(t)$ and $\mathcal{P}_f(t)$, $r_{k,f}^{\mathrm{IDEAL}}(\mathcal{U}_f(t), \mathcal{P}_f(t); t)$, is represented as

$$r_{k,f}^{\mathrm{IDEAL}}\left(\mathcal{U}_f(t), \mathcal{P}_f(t); t\right) = \begin{cases} W \log_2\left(1 + \dfrac{g_{k,f}(t) p_{k,f}(t)}{\displaystyle\sum_{j \in \mathcal{U}_f(t), g_{k,f}(t) < g_{j,f}(t)} g_{k,f}(t) p_{j,f}(t) + 1}\right), & k \in \mathcal{U}_f(t) \\ 0, & k \notin \mathcal{U}_f(t) \end{cases}, \tag{3}$$

where $W$ is the bandwidth of the frequency block.

However, the actual transmission rate, $r_{k,f}^{\mathrm{TX}}(\mathcal{U}_f(t), \mathcal{P}_f(t); t)$, is determined based on the CQI report, $\tilde{g}_{k,f}(t)$, which contains error from $g_{k,f}(t)$ due to the CQI measurement error caused by the noise-plus-interference and CQI reporting delay time in a time-varying channel. If we consider only the initial transmission, the instantaneous throughput of user terminal $k$ at time $t$, $r_k(t)$, becomes

$$r_k(t) = \begin{cases} \sum_{f\in\mathcal{F}} r_{k,f}^{\mathrm{TX}}\left(\mathcal{U}_f(t),\mathcal{P}_f(t);t\right), & \begin{array}{l}\sum_{f\in\mathcal{F}} r_{k,f}^{\mathrm{TX}}\left(\mathcal{U}_f(t),\mathcal{P}_f(t);t\right)\\ \le \sum_{f\in\mathcal{F}} r_{k,f}^{\mathrm{IDEAL}}\left(\mathcal{U}_f(t),\mathcal{P}_f(t);t\right)\end{array} \\ 0, & \text{Otherwise} \end{cases} . \tag{4}$$

Thus, if the sum of the transmission rate, $r_{k,f}^{\mathrm{TX}}(\mathcal{U}_f(t),\mathcal{P}_f(t);t)$, among all frequency blocks is greater than that of the correctly-decodable maximum rate, $r_{k,f}^{\mathrm{IDEAL}}(\mathcal{U}_f(t),\mathcal{P}_f(t);t)$, terminal $k$ fails to decode the received packet and the throughput becomes zero.

To address this problem, we consider reducing the probability of the decoding error by reducing the transmission rate in advance. Here, the effects of decoding errors can differ between eMBB, where decoding error is mainly observed as throughput loss, and URLLC, where decoding error affects the transmission delay time and error rate. Therefore, in the proposed method, the transmission rate, $r_{k,f}^{\mathrm{TX}}(\mathcal{U}_f,\mathcal{P}_f;t)$, is defined as

$$r_{k,f}^{\mathrm{TX}}\left(\mathcal{U}_f,\mathcal{P}_f;t\right) = \begin{cases} W\log_2\left(1+\dfrac{\alpha_{i_k}\tilde{g}_{k,f}(t)p_{k,f}(t)}{\sum\limits_{j\in\mathcal{U}_f(t),\tilde{g}_{k,f}(t)<\tilde{g}_{j,f}(t)}\tilde{g}_{k,f}(t)p_{j,f}(t)+1}\right), & k\in\mathcal{U}_f(t) \\ 0, & k\notin\mathcal{U}_f(t) \end{cases} . \tag{5}$$

By multiplying the predicted value of the signal-to-noise plus interference ratio (SINR) after SIC processing based on the CQI by a service-specific constant, $\alpha_{\mathrm{eMBB}}$ or $\alpha_{\mathrm{URLLC}}$, where $0 \le \alpha_{\mathrm{eMBB}}, \alpha_{\mathrm{URLLC}} \le 1$, the transmission rate is intentionally reduced. As a result, the occurrence of packet decoding errors due to CQI error caused by, e.g., the time fluctuation of the channel can be suppressed. So, the increase in the transmission delay time due to retransmission occurrence can also be reduced. The proposed method assumes that $\alpha_{\mathrm{URLLC}}$ is set lower than $\alpha_{\mathrm{eMBB}}$. This suppresses the occurrence of URLLC packet decoding errors during initial transmission, and aims to achieve both a low error rate and low transmission latency at a high level. In the proposed method, URLLC-prioritized resource allocation is performed using NOMA, and eMBB packets can be multiplexed within the same channel by superposition coding. Therefore, the adverse effects on the bandwidth allocation to eMBB due to such a safe-side transmission rate control in URLLC can be alleviated.

The operation at retransmission in the proposed method is described below. Suppose that a decoding error occurs in the initial transmission packet at time $t_0$ of a certain terminal $k$. The transmission rate of this initial packet is $r_k^{\mathrm{TX}}(t_0) = \sum_{f\in\mathcal{F}} r_{k,f}^{\mathrm{TX}}(\mathcal{U}_f(t_0),\mathcal{P}_f(t_0);t_0)$. Packet retransmission is performed at $t > t_0$, as a result of the resource allocation process for a retransmitted packet for terminal $k$ based on the metric in (2). The resource allocation process for the retransmitted packet is basically the same as that for the initial packet. However, the transmission rate at retransmission, $r_k^{\mathrm{TX}}(t) = \sum_{f\in\mathcal{F}} r_{k,f}^{\mathrm{TX}}(\mathcal{U}_f(t),\mathcal{P}_f(t);t)$, is kept the same as $r_k^{\mathrm{TX}}(t_0)$. Assuming that an appropriate packet combining based on incremental redundancy is conducted among retransmitted packets at the terminal receiver, the throughput of terminal $k$ after a total of $N$ retransmissions that have been performed at $t_1,\ldots,t_N$ is represented as

$$r_k\left(t_N\right) = \begin{cases} r_k^{\mathrm{TX}}\left(t_0\right), & r_k^{\mathrm{TX}}\left(t_0\right) \le \sum_{n=0}^{N}\sum_{f\in\mathcal{F}} r_{k,f}^{\mathrm{IDEAL}}\left(\mathcal{U}_f\left(t_n\right),\mathcal{P}_f\left(t_n\right);t_n\right) \\ 0, & \text{Otherwise} \end{cases} . \tag{6}$$

## 3. Numerical Results

The system-level performance of the proposed method in a cellular environment is evaluated based on computer simulations. Table 1 gives the major simulation parameters. These parameters are based on 5G NR [3]. We assume orthogonal frequency division multiplexing (OFDM) with a 60-kHz subcarrier spacing as the basic signal transmission scheme on which NOMA is actualized based on superposition coding. We assume a 36-MHz system bandwidth with $|\mathcal{F}| = 50$ frequency blocks (resource blocks). The bandwidth, $W$, of the frequency block is 720 kHz, which comprises 12 consecutive subcarriers. Universal frequency reuse is assumed among cells. We assume that there are eMBB and URLLC terminals that are multiplexed within the system bandwidth. The base stations and terminals of each service are placed in random locations within a wrap-around 5×5-square kilometer system coverage area based on the Poisson point process (PPP). The node densities of the base stations and termi-

**Table 1** Simulation parameters.

| | | |
|---|---|---|
| System bandwidth | | 36 MHz |
| Number of frequency blocks | | 50 |
| Coexisting services (use cases) | | eMBB and URLLC |
| Node density | Base station | 1 / km$^2$ |
| | eMBB terminal | 40 / km$^2$ |
| | URLLC terminal | $D_{\mathrm{URLLC}}$ / km$^2$ |
| Transmission power of base station | | 46 dBm |
| Distance-dependent path loss (including antenna gain) | | 114.1+37.6log$_{10}$($r$), $r$: kilometers |
| Shadowing | | Lognormal shadowing with standard deviation of 8 dB and inter-site correlation of 0.5 |
| Instantaneous fading | | Six-path Rayleigh,<br>rms delay spread = 1 μs and $f_{\mathrm{D}}$ = 5.55 Hz |
| Receiver noise power density | | −165 dBm/Hz |
| Traffic model | eMBB terminal | Full buffer |
| | URLLC terminal | 1,000 bits / 10 ms |
| Scheduling interval | | 0.25 ms |
| Minimum interval of retransmission in HARQ | | 1 ms |
| Delay in CQI report | | 10 ms |
| Maximum number of non-orthogonally multiplexed terminals | | 2 per frequency block |

nals receiving eMBB are set to 1 and 40 per square kilometer, respectively. The node density of the terminals receiving URLLC, $D_{\mathrm{URLLC}}$ per square kilometer, is parameterized. The base station transmission power level is 46 dBm. As the propagation model, distance dependent path loss with the decay factor of 3.76 assuming the carrier frequency of 2 GHz; lognormally distributed random shadowing with the standard deviation of 8 dB and inter-site correlation of 0.5; and 6-path Rayleigh fading with the maximum Doppler frequency of 5.55 Hz and the rms delay spread of 1 µs are assumed. The receiver noise power density of the terminal including a noise figure is set to −165 dBm/Hz.

The resource allocation (scheduling and power allocation) interval that equals the packet length is set to 0.25 ms. The maximum number of non-orthogonally multiplexed terminals per frequency block is set to two in the proposed method, which should be sufficient to obtain the most from the potential gain of NOMA based on [13] and [22]. However, the proposed method includes OMA, and if $|\mathcal{U}_f(t)|$ is set to one based on the evaluation results of (2), this means that frequency block $f$ is operated by OMA at time $t$. When NOMA is applied, the fixed power allocation method [22] is used to allocate 80% of the power to terminals experiencing poor channel conditions and the remaining 20% of the power is allocated to terminals experiencing good channel conditions. We also evaluate the case where only OMA is used for comparison. The minimum retransmission interval is 1 ms [3]. The reporting delay of CQI, which is used for transmission rate control, is set to 10 ms.

The traffic model of the eMBB terminal is a full buffer, and the 100-ms average terminal throughput is measured. The system throughput defined by the geometric mean terminal throughput, which corresponds to the proportional fair criterion, is used as the main performance measure of eMBB. On the other hand, we assume that fixed 1,000 information bits are generated every 10 ms to be transmitted to the respective URLLC terminals. We also assume that the information bits that have not been transmitted correctly are discarded 10 ms after the occurrence. The main performance measures for URLLC are the transmission delay time for the correct decoding of a 1,000-bit URLLC payload, and the outage probability of the correct decoding of the 1,000-bit URLLC payload during the required transmission delay time. Parameter $\delta_{\mathrm{eMBB}}$ in the proposed method is fixed at 1.0, and $\delta_{\mathrm{URLLC}}$, $\alpha_{\mathrm{eMBB}}$, and $\alpha_{\mathrm{URLLC}}$ are evaluated as parameters.

Figure 2 shows the cumulative distribution of the average throughput of eMBB terminals with $\alpha_{\mathrm{eMBB}}$ as a parameter. Figure 3 shows the system throughput for eMBB as a function of $\alpha_{\mathrm{eMBB}}$. In these evaluations, an environment where only eMBB user terminals exist with the density of 60 per square kilometer is assumed. The average throughput distribution of eMBB terminals improves when $\alpha_{\mathrm{eMBB}}$ is set to less than one. This is because the decoding error caused by the excessively high transmission rate due to the CQI error can be reduced. However, if $\alpha_{\mathrm{eMBB}}$ is set excessively low, the spectrum efficiency becomes too low and the throughput deteriorates. From Figs. 2 and 3, $\alpha_{\mathrm{eMBB}}$ of 0.7 is the best under the assumed simulation conditions, so in the subsequent evaluations, $\alpha_{\mathrm{eMBB}}$ is set to 0.7.

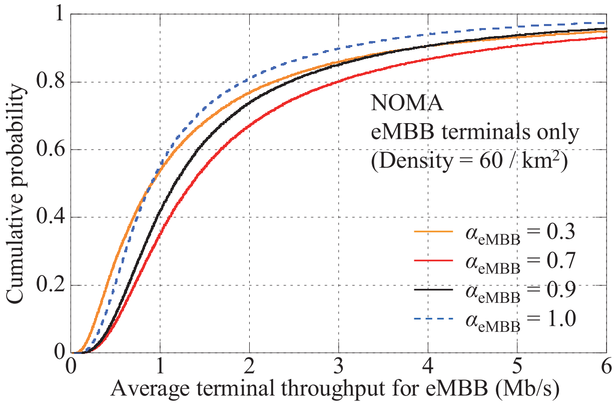


**Fig. 2** Distribution of average terminal throughput for eMBB with $\alpha_{\mathrm{eMBB}}$ as a parameter.

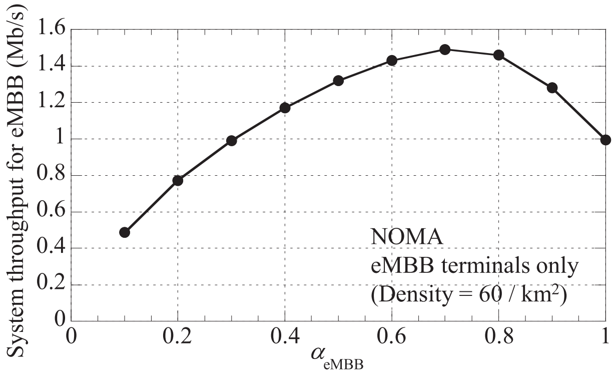


**Fig. 3** System throughput for eMBB as a function of $\alpha_{\mathrm{eMBB}}$.

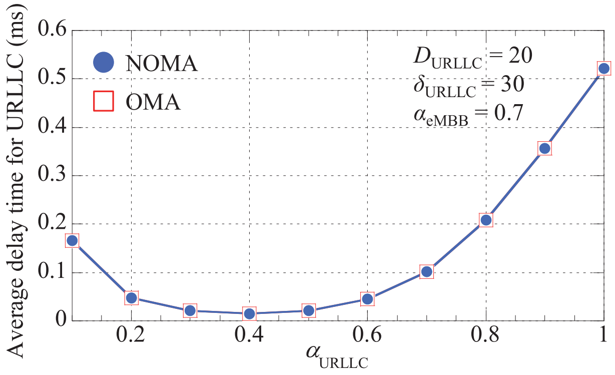


**Fig. 4** Average delay time for URLLC as a function of $\alpha_{\mathrm{URLLC}}$.

Figure 4 shows the average delay time for URLLC as a function of $\alpha_{\mathrm{URLLC}}$. The transmission delay time is defined as the time required to complete correct decoding of all the 1,000-bit URLLC payloads during each 10-ms time interval. The average transmission delay time is calculated for the case where 1,000 information bits could be transmitted correctly within 10 ms. Figure 5 shows the system throughput for eMBB as a function of $\alpha_{\mathrm{URLLC}}$. In Figs. 4 and 5,

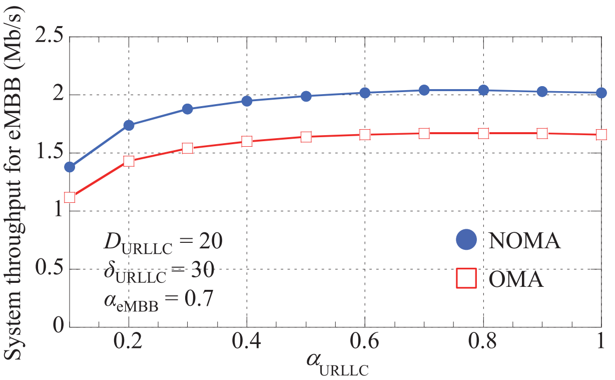


**Fig. 5** System throughput for eMBB as a function of $\alpha_{\mathrm{URLLC}}$.

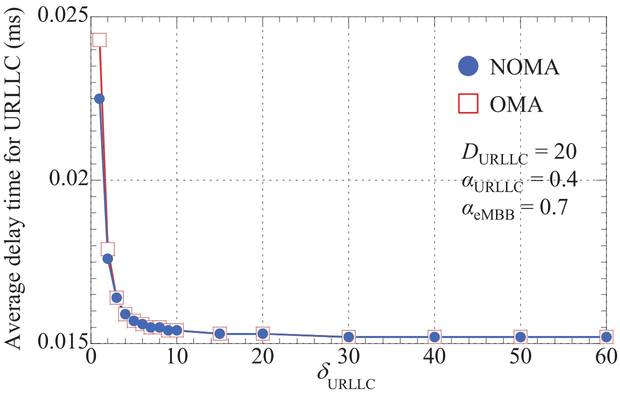


**Fig. 6** Average delay time for URLLC as a function of $\delta_{\mathrm{URLLC}}$.

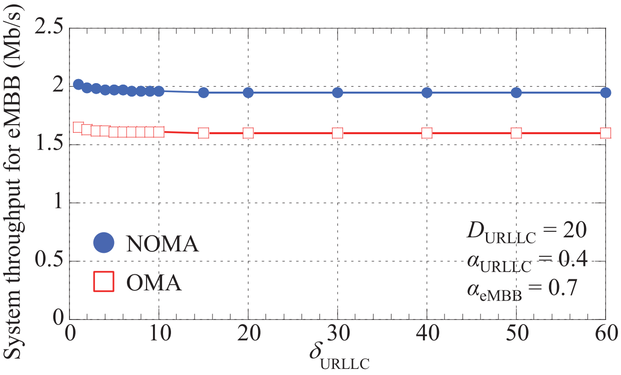


**Fig. 7** System throughput for eMBB as a function of $\delta_{\mathrm{URLLC}}$.

$D_{\mathrm{URLLC}}$ is 20 and $\delta_{\mathrm{URLLC}}$ is set to 30. The proposed NOMA and conventional OMA methods are tested.

Figure 4 shows that the average transmission delay time of URLLC is minimized when $\alpha_{\mathrm{URLLC}}$ is set to 0.4. The average delay time for URLLC is very similar between NOMA and OMA. The increase in the transmission delay time when using an $\alpha_{\mathrm{URLLC}}$ value greater than 0.4 is due to frequent packet retransmissions. This is a result of an excessively high transmission rate due to CQI errors. The incidence of this phenomenon is not different between NOMA and OMA, which may be one reason for nearly the same average delay time for NOMA and OMA. The increase in the transmission delay time when using an $\alpha_{\mathrm{URLLC}}$ value of less than 0.4 is due to the decrease in the number of information bits that can be transmitted per resource allocation interval of 0.25 ms. This increase in delay time is significant for the URLLC user terminal experiencing very poor channel conditions, and 1000 bits cannot be sent even if the entire system bandwidth is allocated to that terminal. Such user terminals experiencing poor channel conditions are allocated high transmission power in NOMA, while the interference from non-orthogonally multiplexed packets of user terminals near the base station cannot be removed by the SIC (this is because the decoding order of the SIC is in the order of user terminals experiencing poor channel conditions). However, for user terminals experiencing very poor channel conditions, the received signal power from the serving base station is low, while the inter-cell interference power is very high, so the effect of interference power from the serving base station due to non-orthogonal packet multiplexing in NOMA is very small. Based on the reasons above, the average transmission delay time for URLLC when the proposed NOMA is used is almost the same as that for OMA.

Figure 5 shows that the system throughput for eMBB is increased using the proposed NOMA method compared to that using OMA. The proposed NOMA method allows prioritized resource allocation to URLLC and achieves the same low transmission latency characteristics as OMA, while the allocated bandwidth to eMBB is increased through non-orthogonal multiplexing. The cost is inter-packet interference caused by non-orthogonal multiplexing. However, this can be effectively suppressed using the SIC. Observing the effect of $\alpha_{\mathrm{URLLC}}$ on the system throughput for eMBB, the use of an excessively low $\alpha_{\mathrm{URLLC}}$ along with the URLLC-prioritized resource allocation using $\delta_{\mathrm{URLLC}}$ of 30 results in system throughput degradation for eMBB since the bandwidth allocation to eMBB terminals is severely limited. On the other hand, when $\alpha_{\mathrm{URLLC}}$ is set excessively high, the bandwidth allocated to URLLC will eventually increase due to the frequent occurrence of retransmission, leading to slight deterioration in the system throughput for eMBB. Considering comprehensively the average transmission delay time of URLLC and the system throughput of eMBB from Figs. 4 and 5, the $\alpha_{\mathrm{URLLC}}$ of approximately 0.4 is an appropriate choice under the assumed simulation conditions.

Figures 6 and 7 show the average transmission delay time for URLLC and the system throughput for eMBB as a function of $\delta_{\mathrm{URLLC}}$, respectively. Density $D_{\mathrm{URLLC}}$ is set to 20, and $\alpha_{\mathrm{URLLC}}$ and $\alpha_{\mathrm{eMBB}}$ are set to 0.4 and 0.7, respectively. Increasing $\delta_{\mathrm{URLLC}}$ from 1.0 promotes preferential resource allocation to URLLC significantly reducing the average transmission delay time for URLLC at the expense of a slight decrease in the eMBB system throughput.

Figures 8 and 9 show the average transmission delay time for URLLC and the system throughput for eMBB as a function of $D_{\mathrm{URLLC}}$, respectively. Parameters $\alpha_{\mathrm{URLLC}}$ and $\alpha_{\mathrm{eMBB}}$ are set to 0.4 and 0.7, respectively. From these figures, we can see that the proposed method suppresses the in-

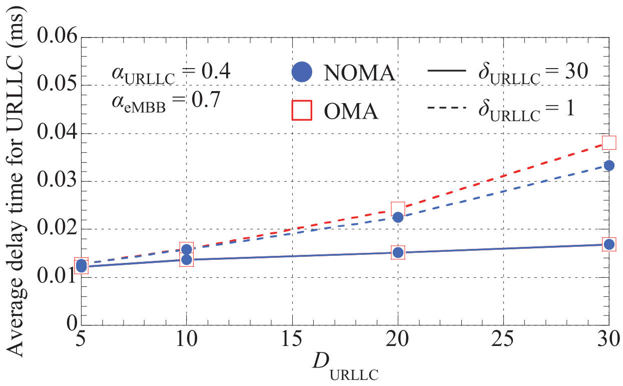


**Fig. 8** Average delay time for URLLC as a function of $D_{\text{URLLC}}$.

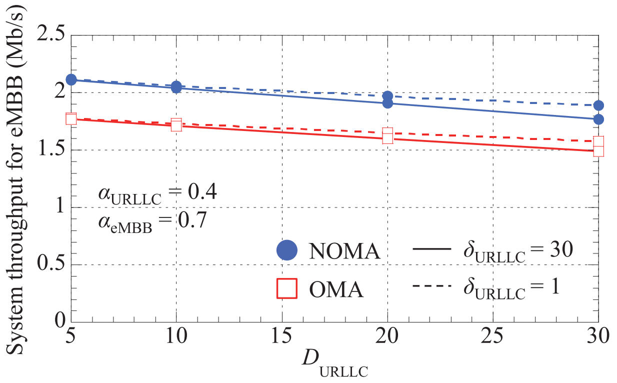


**Fig. 9** System throughput for eMBB as a function of $D_{\text{URLLC}}$.

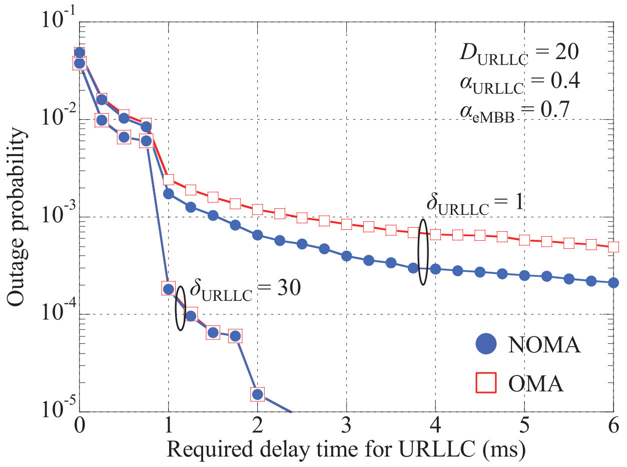


**Fig. 10** Outage probability of required delay time for URLLC.

crease in transmission delay time for URLLC when $D_{\text{URLLC}}$ is increased at the cost of a slight deterioration in the system throughput for eMBB.

Figure 10 shows the outage probability that a 1,000-bit URLLC payload transmission cannot be completed for the required transmission delay time of URLLC, which is indicated on the horizontal axis. Here, $D_{\text{URLLC}}$ is set to 20, and $\alpha_{\text{URLLC}}$ and $\alpha_{\text{eMBB}}$ are set to 0.4 and 0.7, respectively. In the proposed method using NOMA, by setting $\delta_{\text{URLLC}}$ to 30, a very low URLLC outage probability is achieved, while the system throughput of eMBB is significantly increased compared to OMA as shown in Figs. 5 and 7.

In the performance evaluation in the paper, the system bandwidth is fixed at 36 MHz. We note that in the preliminary work in [21], the evaluation of the proposed method assuming a 9-MHz system bandwidth is shown. From the evaluation, we find that the optimal $\alpha_{\text{URLLC}}$ is decreased as the system bandwidth increases. This is because the demerit of using an excessively low $\alpha_{\text{URLLC}}$ in terms of the transmission delay time of URLLC is alleviated with a wider system bandwidth. Meanwhile, the optimal $\alpha_{\text{eMBB}}$ and $\delta_{\text{URLLC}}$ do not appear to be largely dependent on the system bandwidth. As with many LTE and NR parameters, it is difficult to derive the optimal parameters of the proposed method analytically, so it is expected to be set experimentally in the real system. This is done utilizing simulation and other methods, which depend on operational conditions such as the system bandwidth, CQI reporting cycle, and the data size of the URLLC user terminals.

## 4. Conclusion

In this paper, we proposed a system-level NOMA-based HARQ method that achieves ultra-high quality and low latency for URLLC while simultaneously suppressing the adverse effect on the system throughput for eMBB. The proposed method uses radio resource allocation including scheduling to multiplex the URLLC packets and best-effort eMBB packets in the same channel through superposition coding. In order to obtain multi-user diversity between terminals according to the instantaneous channel states of all terminals, we introduced a proportional fair-based resource allocation metric with service-dependent weighting considering the requirements of the respective services. By using this weighting approach, a low error rate and short transmission delay time are explicitly guaranteed. In the proposed method, the transmission rate at the initial transmission is appropriately controlled for each service in order to suppress the decoding errors caused by the excessively high transmission rate caused by the outdated CQI. Computer simulations revealed quantitatively that the proposed method improves the error rate and transmission delay time of URLLC while suppressing the deterioration in the system throughput for eMBB operated at the same time. We note that the SIC processing delay is not considered in this paper for simplicity. In practice, sequential interference cancellation may increase the decoding delay time in the SIC. To mitigate this, it is conceivable to add some scheduling restrictions to avoid the SIC process at the URLLC user terminals. This can be achieved by non-orthogonal multiplexing of the packet for URLLC user terminals with eMBB user terminals whose channel conditions are better than those for URLLC terminals. We note that, in the NOMA adopted in LTE-Advanced [24], the symbol-level complexity reduced maximum likelihood detection (R-ML) [25], which does not require decoding of other terminal signals, is used instead of the SIC under the assumption of QAM modulation. R-ML can provide quite comparable performance to that of the SIC [26]. With R-ML, the decoding processing time on the receiver side does not increase in NOMA and the performance gain by using the proposed method shown in the paper holds.

## References

[1] E. Dahlman, S. Parkvall, J. Sköld, and P. Beming, 3G Evolution: HSPA and LTE for Mobile Broadband, 2nd ed., Academic Press, 2008.

[2] E. Dahlman, S. Parkvall, and J. Sköld, 4G: LTE/LTE-Advanced for Mobile Broadband, 2nd ed., Academic Press, 2013.

[3] E. Dahlman, S. Parkvall, and J. Sköld, 5G NR: The Next Generation Wireless Access Technology, Academic Press, 2018.

[4] NTT DOCOMO, "White paper: 5G evolution and 6G," Jan. 2023.

[5] D.N. Rowitch and L.B. Milstein, "On the performance of hybrid FEC/ARQ systems using rate compatible punctured turbo (RCPT) codes," IEEE Trans. Commun., vol.48, no.6, pp.948–959, June 2000.

[6] Y. Imamura, D. Muramatsu, Y. Kishiyama, and K. Higuchi, "Low latency hybrid ARQ method using channel state information before channel decoding," Proc. APCC2017, Perth, Australia, Dec. 2017.

[7] K. Taniyama, Y. Kishiyama, and K. Higuchi, "Low latency HARQ method using early retransmission prior to channel decoding with multistage decision," Proc. IEEE VTC2019-Fall, Honolulu, U.S.A., Sept. 2019.

[8] K. Miura, Y. Kishiyama, and K. Higuchi, "Low latency HARQ method using early retransmission before channel decoding based on superposition coding," Proc. ISPACS 2019, Taipei, Taiwan, Dec. 2019.

[9] R. Zhang and L. Hanzo, "Superposition-coding aided multiplexed hybrid HARQ scheme for improved link-layer transmission efficiency," Proc. IEEE ICC2009, Dresden, Germany, June 2009.

[10] F. Takahashi and K. Higuchi, "HARQ for predetermined-rate multicast channel," Proc. IEEE VTC 2010-Spring, Taipei, Taiwan, May 2010.

[11] Y. Hasegawa and K. Higuchi, "Bi-directional signal detection and decoding for hybrid ARQ using superposition coding," Proc. IEEE VTC2011-Fall, San Francisco, U.S.A., Sept. 2011.

[12] F. Nadeem, M. Shirvanimoghaddam, Y. Li, and B. Vucetic, "Non-orthogonal HARQ for URLLC: Design and analysis," IEEE Internet Things J., vol.8, no.24, pp.17596–17610, 2021, doi: 10.1109/JIOT.2021.3081698.

[13] K. Higuchi and A. Benjebbour, "Non-orthogonal multiple access (NOMA) with successive interference cancellation for future radio access," IEICE Trans. Commun., vol.E98-B, no.3, pp.403–414, March 2015.

[14] L. Dai, B. Wang, Y. Yuan, S. Han, I. Chih-Lin, and Z. Wang, "Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends," IEEE Commun. Mag., vol.53, no.9, pp.74–81, Sept. 2015.

[15] A. Anand and G. de Veciana, "Resource allocation and HARQ optimization for URLLC traffic in 5G wireless networks," IEEE J. Sel. Areas Commun., vol.36, no.11, pp.2411–2421, Nov. 2018.

[16] S.E. Elayoubi, P. Brown, M. Deghel, and A. Galindo-Serrano, "Radio resource allocation and retransmission schemes for URLLC over 5G networks," IEEE J. Sel. Areas Commun., vol.37, no.4, pp.896–904, April 2019.

[17] W.R. Ghanem, V. Jamali, Y. Sun, and R. Schober, "Resource allocation for multi-user downlink URLLC-OFDMA systems," Proc. IEEE ICC2019, Shanghai, China, May 2019.

[18] B. Chang, L. Zhang, L. Li, G. Zhao, and Z. Chen, "Optimizing resource allocation in URLLC for real-time wireless control systems," IEEE Trans. Veh. Technol., vol.68, no.9, pp.8916–8927, Sept. 2019.

[19] T. Shikuma, Y. Yuda, and K. Higuchi, "NOMA-based optimal multiplexing for multiple downlink service channels to maximize integrated system throughput," IEICE Trans. Commun., vol.E103-B, no.11, pp.1367–1374, Nov. 2020.

[20] C. Wengerter, J. Ohlhorst, and A.G.E. von Elbwart, "Fairness and throughput analysis for generalized proportional fair frequency scheduling in OFDMA," Proc. IEEE VTC2005-Spring, Stockholm, Sweden, pp.1903–1907, May-June 2005.

[21] R. Kobayashi, Y. Yuda, and K. Higuchi, "NOMA-based highly-efficient low-latency HARQ method for URLLC," Proc. IEEE VTC2021-Fall, Virtual Conference, Sept.-Oct. 2021.

[22] N. Otao, Y. Kishiyama, and K. Higuchi, "Performance of non-orthogonal multiple access with SIC in cellular downlink using proportional fair-based resource allocation," IEICE Trans. Commun., vol.E98-B, no.2, pp.344–351, Feb. 2015.

[23] D. Tse and P. Viswanath, Fundamentals of Wireless Communication, Cambridge University Press, 2005.

[24] 3GPP TR36.859 (V13.0.0), "Study on downlink multiuser superposition transmission (MUST) for LTE (Release 13)," Dec. 2015.

[25] C. Yan, A. Harada, A. Benjebbour, Y. Lan, A. Li, and H. Jiang, "Receiver design for downlink non-orthogonal multiple access (NOMA)," Proc. IEEE VTC2015-Spring, Glasgow, U.K., May 2015.

[26] G. Takita, T. Hara, Y. Yuda, and K. Higuchi, "Repetition-based NOMA-HARQ with adaptive termination for URLLC," Proc. IEEE VTC2022-Fall, London and Beijing, Sept. 2022.

**Ryota Kobayashi** received the B.E. and M.E. degrees from Tokyo University of Science, Noda, Japan in 2021 and 2023, respectively. In 2023, he joined NTT DOCOMO INC.

**Yasuaki Yuda** received the B.E. and M.E. degrees from Tokyo University of Science, Japan in 1997 and 1999, respectively. He received the Ph.D. degree from Tokai University, Japan, in 2014. Since 1999, he has been with the Matsushita Electric Industrial Co., Ltd., Panasonic Corporation and Panasonic Holdings Corporation, Japan. His interests are research and development of wireless communication systems.

**Kenichi Higuchi** received the B.E. degree from Waseda University, Tokyo, Japan, in 1994, and received the Dr.Eng. degree from Tohoku University, Sendai, Japan in 2002. In 1994, he joined NTT Mobile Communications Network, Inc. (now, NTT DOCOMO, INC.). While with NTT DOCOMO, INC., he was engaged in the research and standardization of wireless access technologies for wideband DS-CDMA mobile radio, HSPA, LTE, and broadband wireless packet access technologies for systems beyond IMT-2000. In 2007, he joined the faculty of the Tokyo University of Science and currently holds the position of Professor. His current research interests are in the areas of wireless technologies and mobile communication systems, including advanced multiple access, radio resource allocation, inter-cell interference coordination, multiple-antenna transmission techniques, signal processing such as interference cancellation and turbo equalization, and issues related to heterogeneous networks using small cells. He was a co-recipient of the Best Paper Award of the International Symposium on Wireless Personal Multimedia Communications in 2004 and 2007, the Best Paper Award from the IEICE in 2021, a recipient of the Young Researcher's Award from the IEICE in 2003, the 5th YRP Award in 2007, the Prime Minister Invention Prize in 2010, and the Invention Prize of Commissioner of the Japan Patent Office in 2015. He is a member of the IEEE.