INVITED SURVEY PAPER

Non-orthogonal Multiple Access (NOMA) with Successive Interference Cancellation for Future Radio Access

Kenichi HIGUCHI^{†a)}, Member and Anass BENJEBBOUR^{††}, Senior Member

SUMMARY This paper presents our investigation of non-orthogonal multiple access (NOMA) as a novel and promising power-domain user multiplexing scheme for future radio access. Based on information theory, we can expect that NOMA with a successive interference canceller (SIC) applied to the receiver side will offer a better tradeoff between system efficiency and user fairness than orthogonal multiple access (OMA), which is widely used in 3.9 and 4G mobile communication systems. This improvement becomes especially significant when the channel conditions among the non-orthogonally multiplexed users are significantly different. Thus, NOMA can be expected to efficiently exploit the near-far effect experienced in cellular environments. In this paper, we describe the basic principle of NOMA in both the downlink and uplink and then present our proposed NOMA scheme for the scenario where the base station is equipped with multiple antennas. Simulation results show the potential system-level throughput gains of NOMA relative to OMA.

key words: cellular system, non-orthogonal multiple access, superposition coding, successive interference cancellation

1. Introduction

In order to continue to ensure the sustainability of mobile communication services over the next decade, new technology solutions that can respond to future challenges need to be identified and developed [1]. For future radio access in the 2020-era, significant gains in the system capacity/efficiency and quality of user experience (QoE) are required in view of the anticipated exponential increase in the volume of mobile data traffic, e.g., at least 1000-fold in the 2020s compared to 2010. In cellular mobile communications, the design of radio access technology is one important aspect in improving the system capacity in a cost-effective manner. In particular, the multiple access approach is a key part of radio access technology. In the 3rd generation mobile communication systems such as W-CDMA and cdma2000, direct sequence-code division multiple access (DS-CDMA) is used and the receiver is based on simple single-user detection using the Rake receiver. Orthogonal multiple access (OMA) based on orthogonal frequency division multiple access (OFDMA) or single carrier-frequency division multiple access (SC-FDMA) is adopted in the 3.9 and 4th generation mobile communication systems such as LTE [2] and LTE-

Advanced [3], [4]. OMA is a reasonable choice for achieving good system-level throughput performance in packet-domain services using channel-aware time- and frequency-domain scheduling with simple single-user detection at the receiver. However, further enhancements in the system efficiency and QoE, especially at the cell edge, are required in the future. To accommodate such demands, this paper studies non-orthogonal multiple access (NOMA) as a novel and promising candidate multiple access scheme for further system enhancement.

NOMA exploits the new approach of user multiplexing in the power-domain that was not sufficiently utilized in previous generations. In NOMA, multiple users are multiplexed in the power-domain and de-multiplexed on the receiver side using a successive interference canceller (SIC) [5], [6]. Assuming that the transmit signal is generated based on orthogonal frequency division multiplexing (OFDM) including discrete Fourier transform (DFT)-spread OFDM [2], which is robust against multipath interference, and the channelization is solely obtained through capacity-achieving channel codes such as the turbo code and low-density parity check (LDPC) code, non-orthogonal user multiplexing forms superposition coding [6]. From an information-theoretic perspective, NOMA with a SIC is an optimal multiple access scheme from the viewpoint of the achievable multiuser capacity region, in the downlink [8]–[13] and in the uplink [14], [15]. In fact, the wider the gap between the multiuser capacity regions of NOMA and OMA, the more NOMA can contribute to the simultaneous enhancement of the system efficiency and cell-edge user experience. This is of particular importance in cellular systems where the channel conditions vary significantly among users due to the near-far effect.

Recently, there have been several investigations on advanced schemes for non-orthogonal signal transmission for single user transmission and non-orthogonal user multiplexing among multiple users. For example, new non-orthogonal waveforms such as filter bank multicarrier (FBMC) [16] and universal filtered multicarrier (UFMC) [17] have been investigated for enhancing the spectral efficiency of OFDM signals assuming advanced receivers [18]. Faster-than-Nyquist (FTN) signaling [19] is another approach for non-orthogonal signal transmission within a single user transmission that exploits the excess bandwidth of the signal. Interleaved division multiple access (IDMA) [20], [21], where the channelization of the respective users is achieved by the user-specific channel interleaver and mul-

Manuscript received June 9, 2014.

Manuscript revised October 3, 2014.

[†]The author is 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 NTT DOCOMO, INC., Yokosuka-shi, 239-8536 Japan.

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

DOI: 10.1587/transcom.E98.B.403

tiuser detection at the receiver, is investigated to accommodate large numbers of low-rate users [22]. However, these schemes do not exploit the channel difference among users and generally require high complexity receivers for signal recovery or separation.

This paper presents our investigations on NOMA for both the downlink and uplink [23]–[35]. The remainder of the paper is organized as follows. Section 2 describes the basic principles behind multiuser capacity enhancement achieved by NOMA. Section 3 describes the system model of NOMA with multiple antennas used at the base station. We also present our proposed downlink NOMA scheme, which is characterized by the intra-beam superposition coding and SIC [28]. In Sect. 3, the resource allocation (bandwidth, which includes time, frequency, and beam in the downlink, and power) problem for NOMA is discussed. Section 4 presents the system-level throughput gains of NOMA compared to OMA. Finally, Sect. 5 concludes the paper.

2. Basic Principle

2.1 Downlink

In this subsection, we describe the basic principle of downlink NOMA and the potential gains we expect relative to OMA. For simplicity, we assume here two users and a single transmitter and receiver antenna. The overall system transmission bandwidth is assumed to be one Hertz. Figure 1 shows a spectrum usage comparison between NOMA and OMA in the downlink.

In the downlink, the base station transmits a signal for user i ($i = 1, 2$), s_i , where $E[|s_i|^2] = 1$, with transmission power p_i . The sum of p_i is equal to p_{total} . In NOMA, s_1 and s_2 are superposition coded as

$$x = \sqrt{p_1}s_1 + \sqrt{p_2}s_2. \quad (1)$$

The received signal at user i is represented as

$$y_i = h_i x + w_i, \quad (2)$$

where h_i is the complex channel coefficient between user i and the base station. Term w_i denotes the receiver Gaussian noise including inter-cell interference at the user i receiver.

Fig. 1 Spectrum usage comparison in downlink.

The power spectral density of w_i is $N_{0,i}$.

In downlink NOMA, the SIC process is implemented at the user terminal receiver. The optimal order of SIC decoding in the downlink is different from that in the uplink. The optimal order for decoding in the downlink is in the order of the increasing channel gain normalized by the noise plus inter-cell interference power, $|h_i|^2/N_{0,i}$ [5], [6], [8]–[11]. Based on this order, every user is able to decode correctly and remove interference from the signals of other users whose decoding order in the SIC process comes before that user. This is possible because the transmission rate of user j is controlled so that it can be correctly decoded at the user j receiver, thus it can also be decoded at any other user i with $|h_i|^2/N_{0,i} > |h_j|^2/N_{0,j}$.

More specifically, for a 2-user NOMA case with $|h_1|^2/N_{0,1} > |h_2|^2/N_{0,2}$, user 2 does not perform interference cancellation since it comes first in the decoding order; whereas, user 1 first decodes s_2 and subtracts its component from received signal y_1 before decoding its own signal. Therefore, the throughput of user i , R_i , is represented as

$$\begin{cases} R_1 = \log_2(1 + p_1|h_1|^2/N_{0,1}) \\ R_2 = \log_2(1 + p_2|h_2|^2/(p_1|h_2|^2 + N_{0,2})) \end{cases} \quad (3)$$

On the other hand, in downlink OMA with the bandwidth of β ($0 \leq \beta \leq 1$) Hz assigned to user 1 and the remaining bandwidth, $1-\beta$ Hz, assigned to user 2, R_i is represented as

$$\begin{cases} R_1 = \beta \log_2(1 + p_1|h_1|^2/\beta N_{0,1}) \\ R_2 = (1-\beta) \log_2(1 + p_2|h_2|^2/(1-\beta)N_{0,2}) \end{cases} \quad (4)$$

Figures 2(a) and 2(b) show the throughput (capacity) regions for the cases with symmetric and asymmetric channels, respectively. In the symmetric channel, the signal-to-noise ratios (SNRs), $p_{\text{total}}|h_i|^2/N_{0,i}$, of the two users are

(a) Symmetric channel

(b) Asymmetric channel

Fig. 2 Throughput (capacity) region in downlink.

the same, while they are different in the asymmetric channel. In Fig. 2(a), both $p_{\text{total}}|h_1|^2/N_{0,1}$ and $p_{\text{total}}|h_2|^2/N_{0,2}$ are set to 10 dB. In Fig. 2(b), $p_{\text{total}}|h_1|^2/N_{0,1}$ and $p_{\text{total}}|h_2|^2/N_{0,2}$ are set to 20 and 0 dB, respectively. In the symmetric channel case, the throughput regions for OMA and NOMA with the SIC are identical. In the asymmetric channel case, the maximum total throughput is achieved when all the transmission power (and bandwidth) is allocated to user 1 only, which is achieved by both multiple access schemes. However, the throughput region of NOMA with SIC is much wider than that for OMA in the asymmetric channel case. For example, if we want R_2 to be 0.8 b/s, the achievable R_1 for NOMA with a SIC is approximately 2-fold higher than that for OMA. This is because the throughput of user 1 with a high $p_{\text{total}}|h_1|^2/N_{0,1}$ is bandwidth-limited rather than power-limited and superposition coding with user 2 allows user 1 to use the full bandwidth while being allocated only a small amount of transmission power because of power sharing with user 2. Thus, user 1 imparts only a small amount of interference, $p_1|h_2|^2$, to user 2. In contrast, OMA has to allocate a significant fraction of bandwidth to user 2 to increase its throughput, and this causes severe degradation in the throughput of user 1 whose throughput is bandwidth-limited. Therefore, in the cellular downlink where the channel conditions are different among users due to the near-far effect, NOMA has the potential to improve the tradeoff between the system efficiency and user fairness compared to OMA.

2.2 Uplink

In the uplink, user i transmits signal s_i , with transmission power p_i . Figure 3 shows a spectrum usage comparison between NOMA and OMA in the uplink.

For 2-user uplink NOMA, the received signal at the base station is represented as

$$y = \sqrt{p_1}h_1s_1 + \sqrt{p_2}h_2s_2 + w, \quad (5)$$

where w denotes the receiver Gaussian noise including inter-cell interference. The power spectral density of w is N_0 .

In NOMA, s_1 and s_2 are transmitted using the same frequency and interfere with each other. In the uplink, the SIC is implemented at the base station receiver. With the

Fig. 3 Spectrum usage comparison in uplink.

SIC, the receiver decodes s_1 and s_2 in two stages. In the first stage, the receiver decodes s_1 , treating s_2 as Gaussian interference. Once the receiver correctly decodes s_1 , it can subtract the s_1 component from aggregate received signal y and then decode s_2 . Since there is now only the receiver Gaussian noise remains in the system, the maximum throughput is achieved for user 2 the same as its single-user bound. Thus, the throughput of user i , R_i , is represented as

$$\begin{cases} R_1 = \log_2(1 + p_1|h_1|^2/(p_2|h_2|^2 + N_0)) \\ R_2 = \log_2(1 + p_2|h_2|^2/N_0) \end{cases} \quad (6)$$

If the decoding order is opposite, R_i becomes

$$\begin{cases} R_1 = \log_2(1 + p_1|h_1|^2/N_0) \\ R_2 = \log_2(1 + p_2|h_2|^2/(p_1|h_1|^2 + N_0)) \end{cases} \quad (7)$$

Note that in both (6) and (7), regardless of the decoding order in the SIC, the total throughput, $R_1 + R_2$, is the same as $\log_2(1 + (p_1|h_1|^2 + p_2|h_2|^2)/N_0)$, which is indeed the maximum total achievable throughput in this uplink multiple-access channel.

On the other hand, for uplink OMA with the bandwidth of β Hz assigned to user 1 and the remaining bandwidth, $1 - \beta$ Hz assigned to user 2, R_i is represented as

$$\begin{cases} R_1 = \beta \log_2(1 + p_1|h_1|^2/\beta N_0) \\ R_2 = (1 - \beta) \log_2(1 + p_2|h_2|^2/(1 - \beta)N_0) \end{cases} \quad (8)$$

Figures 4(a) and 4(b) show the throughput (capacity) regions for the cases with symmetric and asymmetric channels, respectively. In Fig. 4(a), both $p_1|h_1|^2/N_0$ and $p_2|h_2|^2/N_0$ are set to 10 dB. In Fig. 4(b), $p_1|h_1|^2/N_0$ and $p_2|h_2|^2/N_0$ are set to 20 and 0 dB, respectively. NOMA with the SIC achieves Point A by decoding the signal for user 1 first. Point B is achieved when the signal for user 2 is decoded first. The throughput pair on line segment AB achieves the maximum total throughput. In the symmetric channel, although the throughput region of OMA is strictly

(a) Symmetric channel

(b) Asymmetric channel

Fig. 4 Throughput (capacity) region in uplink.

a subset of that for NOMA with the SIC, OMA achieves the maximum total throughput at a single point. At that point, the throughput of the two users is equal. However, in the asymmetric channel, although OMA achieves the maximum total throughput, the user throughput is very unfair. Throughput R_2 is 0.067 b/s using OMA, which is approximately 1/15 that of the single-user bound. This is because OMA must allocate most of the bandwidth exclusively to the user under good channel conditions to achieve the maximum total throughput. If we want to achieve the R_2 of 0.8 b/s to improve the user fairness in OMA, R_1 is severely degraded to approximately 3.70 b/s due to the largely reduced bandwidth assignment. On the other hand, NOMA with the SIC achieves the R_2 of 1.0, which is the single-user bound while achieving the maximum total throughput (at Point A).

Given the above analysis, we expect that in a cellular uplink as well as the downlink, NOMA has the potential to yield a better tradeoff between the system efficiency and user fairness compared to OMA.

3. NOMA with Multiple Antennas at Base Station

This section describes the system model of the proposed NOMA scheme with multiple antennas used at the base station. The resource allocation (bandwidth, which includes time, frequency, and beam in the downlink, and power) scheme for NOMA is also presented. Figure 5 summarizes the component technologies of NOMA discussed in this section.

3.1 Downlink

Here, we describe the proposed downlink NOMA scheme with multiple transmit antennas used at the base station. In single-input single-output (SISO) and single-input multiple-output (SIMO) downlink, superposition coding with a SIC receiver and transmitter dirty paper coding (DPC) [7] are equivalent from the viewpoint of the achievable multiuser capacity region. However, in the multiple-input multiple-output (MIMO) downlink, the downlink channel (i.e., broadcast channel) is not degraded. Therefore, superposition coding with a SIC is not optimal and the DPC

should be used to achieve the entire multiuser capacity region [12]. However, DPC is difficult to implement in practice and is very sensitive to delay in the feedback of the channel state information to the base station transmitter. Furthermore, in order to achieve the multiuser capacity region using DPC, a user-dependent beamforming (precoding) must be employed. This results in increased overhead of the (orthogonal) reference signals dedicated to the respective users as the number of multiplexed users is increased beyond the number of transmitter antennas. This increase in overhead of the reference signal decreases the achievable throughput gain from the DPC in practice.

Therefore, we proposed a downlink NOMA scheme using intra-beam superposition coding and SIC [28]. In the proposed NOMA scheme, the number of transmitter beams is restricted to the number of transmitter antennas, which is the same as in OMA in LTE-Advanced. Within each beam, multiple user signals are superposition coded. This is called intra-beam superposition coding. With this non-orthogonal user multiplexing scheme, the number of reference signals is equal to the number of transmitter antennas, which is the same as in LTE-Advanced, irrespective of the number of non-orthogonally multiplexed users. Furthermore, at the user terminal, the inter-beam interference is suppressed by spatial filtering using multiple receiver antennas. After spatial filtering, the channel of the non-orthogonally multiplexed users by superposition coding within a beam becomes degraded. Thus, successive interference cancellation is carried out on the spatially-filtered scalar received signal to remove the inter-user interference within a beam due to intra-beam superposition coding. The proposed NOMA scheme is much easier to implement and more robust against channel variations compared to DPC since both spatial filtering and the SIC rely on channel state information available on the receiver side.

In the following, the system model of the proposed downlink NOMA scheme is described. We assume OFDM signaling with a cyclic prefix. Therefore, the inter-symbol interference and inter-carrier interference are perfectly eliminated assuming that the length of the cyclic prefix is sufficiently long so that it covers the entire multipath delay spread. There are F frequency blocks and the bandwidth of a frequency block is W Hz. The number of transmitter antennas at the base station is M . The number of receiver antennas at the user terminal is N . The number of users per cell is K . For simplicity, in the following, we describe the proposed scheme at some particular time-frequency block (resource block) f ($f = 1, \dots, F$). For multiple time-frequency blocks, the same process is performed independently in principle. In this section, the time index, t , is omitted for simplicity.

The base station performs MIMO transmission with B beams, where $1 \leq B \leq M$. The M -dimensional b -th ($b = 1, \dots, B$) transmitter beamforming (precoding) vector at frequency block f is denoted as $\mathbf{m}_{f,b}$. We assume that the multiuser scheduler schedules a set of users, $U_{f,b} = \{i(f, b, 1), i(f, b, 2), \dots, i(f, b, k(f, b))\}$, to beam b of fre-

Fig. 5 Component technologies of NOMA.

quency block f . Term $i(f, b, u)$ indicates the u -th ($u = 1, \dots, k(f, b)$) user index scheduled at beam b of frequency block f , and $k(f, b)$ ($k(f, b) \leq K$) denotes the number of simultaneously scheduled users at beam b of frequency block f . At the base station transmitter, each $i(f, b, u)$ -th user information bit sequence is independently channel coded and modulated. Term $s_{i(f, b, u), f, b}$ denotes the coded modulation symbol of user $i(f, b, u)$ at beam b of a certain subcarrier of frequency block f . We assume $E[|s_{i(f, b, u), f, b}|^2] = 1$. The allocated transmission power for user $i(f, b, u)$ at beam b of frequency block f is denoted as $p_{i(f, b, u), f, b}$. In the proposed scheme, $s_{i(f, b, u), f, b}$ of all $k(f, b)$ users is first superposition coded as intra-beam superposition coding and then multiplied by the transmitter beamforming vector, $\mathbf{m}_{f, b}$. Finally, by accumulating all B beam transmission signal vectors, the M -dimensional transmission signal vector, $\mathbf{x}_f$, at frequency block f is generated as

$$\mathbf{x}_f = \sum_{b=1}^B \mathbf{m}_{f, b} \sum_{u=1}^{k(f, b)} \sqrt{p_{i(f, b, u), f, b}} s_{i(f, b, u), f, b}. \quad (9)$$

The transmission power allocation constraint is represented as

$$\sum_{u=1}^{k(f, b)} p_{i(f, b, u), f, b} = p_b, \quad \sum_{b=1}^B p_b = p_{\text{total}}, \quad (10)$$

where p_b is the transmission power of beam b and p_{total} is the total transmission power. We assume that p_b and p_{total} are identical for all frequency blocks in the paper. The set of $p_{i(f, b, u), f, b}$ at beam b of frequency block f is denoted as $P_{f, b}$. Figure 6 illustrates the transmission signal generation in the proposed NOMA scheme at frequency block f .

The N -dimensional received signal vector of user $i(f, b, u)$ at frequency block f , $\mathbf{y}_{i(f, b, u), f}$, is represented as

$$\begin{aligned} \mathbf{y}_{i(f, b, u), f} &= \mathbf{H}_{i(f, b, u), f} \mathbf{x}_f + \mathbf{w}_{i(f, b, u), f} \\ &= \mathbf{H}_{i(f, b, u), f} \sum_{b'=1}^B \mathbf{m}_{f, b'} \sum_{u'=1}^{k(f, b')} \sqrt{p_{i(f, b', u'), f, b'}} s_{i(f, b', u'), f, b'} \\ &\quad + \mathbf{w}_{i(f, b, u), f}, \end{aligned} \quad (11)$$

Fig. 6 Downlink transmission signal generation in proposed NOMA scheme (at frequency block f).

where $\mathbf{H}_{i(f, b, u), f}$ is the $N \times M$ -dimensional channel matrix between the base station and user $i(f, b, u)$ at frequency block f and $\mathbf{w}_{i(f, b, u), f}$ denotes the receiver noise plus inter-cell interference vector at frequency block f .

In the proposed scheme, the user terminal first performs spatial filtering to suppress the inter-beam interference. Assuming that user $i(f, b, u)$ uses the N -dimensional spatial filtering vector, $\mathbf{v}_{i(f, b, u), f, b}$, to receive beam b of frequency block f , the scalar signal after the spatial filtering, $z_{i(f, b, u), f, b}$, is represented as

$$\begin{aligned} z_{i(f, b, u), f, b} &= \mathbf{v}_{i(f, b, u), f, b}^H \mathbf{y}_{i(f, b, u), f} \\ &= \mathbf{v}_{i(f, b, u), f, b}^H \mathbf{H}_{i(f, b, u), f} \mathbf{m}_{f, b} \sum_{u'=1}^{k(f, b)} \sqrt{p_{i(f, b, u'), f, b}} s_{i(f, b, u'), f, b} \\ &\quad + \mathbf{v}_{i(f, b, u), f, b}^H \mathbf{H}_{i(f, b, u), f} \sum_{b'=1, b' \neq b}^B \mathbf{m}_{f, b'} \sum_{u'=1}^{k(f, b')} \sqrt{p_{i(f, b', u'), f, b'}} s_{i(f, b', u'), f, b'} \\ &\quad + \mathbf{v}_{i(f, b, u), f, b}^H \mathbf{w}_{i(f, b, u), f}, \end{aligned} \quad (12)$$

In the paper, we assume that $\mathbf{v}_{i(f, b, u), f, b}$ is calculated based on the minimum mean squared error (MMSE) criterion. The second and third terms of (12) are the inter-beam interference and receiver noise plus inter-cell interference observed at the spatial filtering output, respectively. By normalizing the aggregated power of the inter-beam interference and receiver noise plus inter-cell interference as one, (12) can be rewritten as

$$z_{i(f, b, u), f, b} = \sqrt{g_{i(f, b, u), f, b}} \sum_{u'=1}^{k(f, b)} \sqrt{p_{i(f, b, u'), f, b}} s_{i(f, b, u'), f, b} + q_{i(f, b, u), f, b}, \quad (13)$$

where $q_{i(f, b, u), f, b}$ denotes the sum of the inter-beam interference, receiver noise, and inter-cell interference terms after normalization (i.e., $E[|q_{i(f, b, u), f, b}|^2] = 1$). Term $g_{i(f, b, u), f, b}$ is represented as

$$g_{i(f, b, u), f, b} = \frac{|\mathbf{v}_{i(f, b, u), f, b}^H \mathbf{H}_{i(f, b, u), f} \mathbf{m}_{f, b}|^2}{\left\{ \sum_{b'=1, b' \neq b}^B p_{b'} |\mathbf{v}_{i(f, b, u), f, b}^H \mathbf{H}_{i(f, b, u), f} \mathbf{m}_{f, b'}|^2 + \mathbf{v}_{i(f, b, u), f, b}^H E[\mathbf{w}_{i(f, b, u), f} \mathbf{w}_{i(f, b, u), f}^H] \mathbf{v}_{i(f, b, u), f, b} \right\}}. \quad (14)$$

Thus, for the users to which beam b of frequency block f is allocated, the channel after the spatial filtering can be regarded as a degraded SISO channel, and the normalized equivalent channel gain of user $i(f, b, u)$ becomes $g_{i(f, b, u), f, b}$.

We apply the intra-beam SIC to signal $z_{i(f, b, u), f, b}$ in order to remove the inter-user interference within a beam. Similar to the SISO downlink [5], [6], [9], the optimal order of decoding is in the order of the increasing normalized channel gain, $g_{i(f, b, u), f, b}$. Based on this order, any user can correctly decode and cancel the signals of other users whose decoding order comes before that user. Thus, user $i(f, b, u)$ can remove the inter-user interference from user $i(f, b, u')$

Fig. 7 Operational principle of proposed downlink NOMA scheme.

whose $g_{i(f,b,u'),f,b}$ is lower than $g_{i(f,b,u),f,b}$. As a result, the throughput of user k ($k = 1, \dots, K$) at beam b of frequency block f assuming the scheduled user set $U_{f,b}$ and allocated transmission power set $P_{f,b}$ is represented as

$$R_{f,b}(k|U_{f,b}, P_{f,b}) = \begin{cases} W \log_2 \left(1 + \frac{g_{k,f,b} P_{k,f,b}}{\sum_{j \in U_{f,b}, g_{k,f,b} < g_{j,f,b}} g_{j,f,b} P_{j,f,b} + 1} \right), & k \in U_{f,b} \\ 0, & k \notin U_{f,b} \end{cases} \quad (15)$$

In the actual transmission procedure, before the transmission, the base station sets the transmission rate of the respective user signals, $s_{i(f,b,u),f,b}$, using (15) based on the channel state information of all users so that the achievable maximum throughput is obtained for the respective users.

Figure 7 illustrates the operational principle of the proposed scheme using intra-beam superposition coding at the base station transmitter and intra-beam SIC at the user terminal receiver.

3.2 Uplink

Here, we describe the system model for the uplink NOMA scheme, assuming M base station antennas. For the sake of simplicity, we assume a single transmitter antenna per user. However, the following description can be easily extended to the case with multiple transmitter antennas per user.

We assume that the multiuser scheduler schedules a set of users, $U_f = \{i(f, 1), i(f, 2), \dots, i(f, k(f))\}$, to frequency block f , where $i(f, u)$ indicates the u -th ($u = 1, \dots, k(f)$) user index scheduled at frequency block f and $k(f)$ ($k(f) \leq K$) denotes the number of scheduled users at frequency block f . At the $i(f, u)$ -th user terminal transmitter, coded modulation symbol $s_{i(f,u),f}$ ($E[|s_{i(f,u),f}|^2] = 1$) is transmitted on a certain subcarrier of frequency block f . The M -dimensional received signal vector of frequency block f , $\mathbf{y}_f$, at the base station is represented as

$$\mathbf{y}_f = \sum_{u=1}^{k(f)} \mathbf{h}_{i(f,u),f} \sqrt{p_{i(f,u),f}} s_{i(f,u),f} + \mathbf{w}_f, \quad (16)$$

where $\mathbf{h}_{i(f,u),f}$ is the M -dimensional channel coefficient vec-

Fig. 8 MMSE-SIC.

tor between user $i(f, u)$ and the serving base station at frequency block f . Term $p_{i(f,u),f}$ is the transmission power of user $i(f, u)$ at frequency block f . The set of $p_{i(f,u),f}$ at frequency block f is denoted as P_f . The M -dimensional vector, $\mathbf{w}_f$, is the receiver noise and inter-cell interference vector. To simplify, we assume that $\mathbf{w}_f$ consists of i.i.d. circularly symmetric complex Gaussian random variables with zero mean and the variance of $N_{0,f}$. Given the received signal vector in (16), the maximum of the total user throughput, $R_{\text{total},f}(U_f, P_f)$, is represented as

$$R_{\text{total},f}(U_f, P_f) = W \log \det \left(\mathbf{I} + \frac{\sum_{u=1}^{k(f)} p_{i(f,u),f} \mathbf{h}_{i(f,u),f} \mathbf{h}_{i(f,u),f}^H}{N_{0,f}} \right). \quad (17)$$

The receiver structure that achieves $R_{\text{total},f}$ in (17) is the MMSE-based linear filtering followed by a SIC (MMSE-SIC) [5]. Figure 8 shows the MMSE-SIC receiver assuming that the order of decoding is $\pi(f, 1), \pi(f, 2), \dots, \pi(f, k(f))$, where $\pi(f)$ represents an arbitrary permutation of $\{i(f, u)\}$.

$$\pi(f) = \begin{pmatrix} \pi(f, 1) & \pi(f, 2) & \cdots & \pi(f, k(f)) \\ i(f, 1) & i(f, 2) & \cdots & i(f, k(f)) \end{pmatrix}. \quad (18)$$

In the MMSE-SIC, the input of the MMSE filtering for signal detection of user $\pi(f, u)$ is the received signal vector after subtraction of the signal component from user $\pi(f, 1)$ up to $\pi(f, u - 1)$, via the SIC process. Thus, the input for the MMSE filter for signal detection of user $\pi(f, u)$ is represented as

$$\mathbf{h}_{\pi(f,u),f} \sqrt{p_{\pi(f,u),f}} s_{\pi(f,u),f} + \sum_{u'=u+1}^{k(f)} \mathbf{h}_{\pi(f,u'),f} \sqrt{p_{\pi(f,u'),f}} s_{\pi(f,u'),f} + \mathbf{w}_f. \quad (19)$$

The MMSE filter vector for user $\pi(f, u)$, $\mathbf{c}_{\pi(f,u),f}$, is represented as

$$\mathbf{c}_{\pi(f,u),f} = \left(\sum_{u'=u}^{k(f)} p_{\pi(f,u'),f} \mathbf{h}_{\pi(f,u'),f} \mathbf{h}_{\pi(f,u'),f}^H + N_{0,f} \mathbf{I} \right)^{-1} \cdot \sqrt{p_{\pi(f,u),f}} \mathbf{h}_{\pi(f,u),f}. \quad (20)$$

The signal-to-interference and noise power ratio (SINR) of $s_{\pi(f,u),f}$ at the MMSE-filter output is represented as

$$\text{SINR}_{\pi(f,u),f} = \frac{\sqrt{p_{\pi(f,u),f}} \mathbf{c}_{\pi(f,u),f}^H \mathbf{h}_{\pi(f,u),f}}{1 - \sqrt{p_{\pi(f,u),f}} \mathbf{c}_{\pi(f,u),f}^H \mathbf{h}_{\pi(f,u),f}}. \quad (21)$$

Therefore, the throughput of user k at frequency block f assuming the scheduled user set of U_f , transmission power set P_f , and decoding order $\pi(f)$ is represented as

$$R_f(k|U_f, P_f, \pi_f) = \begin{cases} W \log(1 + \text{SINR}_{\pi(f,u),f}) \\ = -W \log \left(\frac{1 - p_{\pi(f,u),f} \mathbf{h}_{\pi(f,u),f}^H \mathbf{h}_{\pi(f,u),f}}{\sum_{u'=u}^{(k,f)} p_{\pi(f,u'),f} \mathbf{h}_{\pi(f,u'),f} \mathbf{h}_{\pi(f,u'),f}^H + N_{0,f} \mathbf{I}} \right)^{-1} \mathbf{h}_{\pi(f,u),f} \\ = -W \log \det \left(\mathbf{I} - \left(\sum_{u'=u}^{(k,f)} p_{\pi(f,u'),f} \mathbf{h}_{\pi(f,u'),f} \mathbf{h}_{\pi(f,u'),f}^H + N_{0,f} \mathbf{I} \right)^{-1} \right) \\ = W \log \det \left(\mathbf{I} + \frac{\sum_{u'=u}^{(k,f)} p_{\pi(f,u'),f} \mathbf{h}_{\pi(f,u'),f} \mathbf{h}_{\pi(f,u'),f}^H}{N_{0,f}} \right) \\ -W \log \det \left(\mathbf{I} + \frac{\sum_{u'=u+1}^{(k,f)} p_{\pi(f,u'),f} \mathbf{h}_{\pi(f,u'),f} \mathbf{h}_{\pi(f,u'),f}^H}{N_{0,f}} \right), k = \pi(f, u) \in U_f \\ 0, k \notin U_f \end{cases} \quad (22)$$

We see that $\sum_u R_f(\pi(f, u)|U_f, P_f, \pi_f)$ equals $R_{\text{total},f}(U_f, P_f)$ for any π_f . Since any user can benefit from increased throughput when decoded at a later stage of the SIC process, we assume that π_f is determined so that the user within U_f is sorted in the order of the decreasing average user throughput in order to improve user fairness. This ordering has an additional merit when the channel coding block is distributed over multiple frequency blocks, as in LTE and LTE-Advanced for example [2]–[4]. Since the ordering is determined by the average user throughput, which is not dependent on the frequency block index, the order between any combination of users i and j in the cell becomes common in all frequency blocks. This property is essential for the SIC process when the channel coding block is distributed over multiple frequency blocks.

In the actual transmission procedure, before the transmission, the base station determines the transmission rate of the respective user signals, $s_{i(f,u),f}$, using (22) based on the channel state information of all users so that the achievable maximum throughput is obtained for the respective users. The respective users are informed of the determined transmission rate along with the assigned resource information using the downlink control signaling.

3.3 Resource Allocation

The radio resource allocation policy significantly affects the system efficiency (measured by, for example, the total average user throughput) and cell-edge user experience (measured by, for example, the cell-edge average user throughput). In this section, we describe the radio resource allocation for NOMA in both the downlink and uplink. For the uplink, the beam index is omitted in the following description.

The average user throughput of user k assuming

scheduling user set $U_{f,b}$ and power set $P_{f,b}$ at time t is defined as

$$T(k; t) = T(k; t-1) + \frac{1}{t_c} \left\{ \sum_{f=1}^F \sum_{b=1}^B R_{f,b}(k|U_{f,b}, P_{f,b}; t) - T(k; t-1) \right\}. \quad (23)$$

Term t denotes the time index representing a subframe index. Parameter t_c defines the time window used for throughput averaging. In the evaluations, we assume a t_c of 100 with a subframe length of 1 ms. Thus, the 100-ms average user throughput is measured. Term $R_{f,b}(k|U_{f,b}, P_{f,b}; t)$ is the throughput of user k in beam b of frequency block b at time instance t assuming $U_{f,b}$ and $P_{f,b}$, which is calculated using (15) and (22) for the downlink and uplink, respectively.

Let $F(t)$ be the objective function of resource allocation to be maximized. In this paper, we use the (p, α) -proportionally fairness criterion [36], [37]. Thus, $F(t)$ is represented as

$$F(t) = \sum_{k=1}^K \rho_k \frac{T^{1-\alpha}(k; t)}{1-\alpha}, \quad (24)$$

where ρ_k ($\rho_k > 0$) denotes the per-user weight and α ($\alpha \geq 0$) is the control parameter, which controls the per-user fairness with regard to the throughput performance. When ρ_k for all users is the same, the α of one corresponds to pure proportional fair resource allocation [38], [39], while the α of 0 and ∞ correspond to the total user throughput maximization criterion [40] and Max-Min user throughput criterion [41], respectively. Note that the greater the value of α is the better the user fairness is but at the cost of reduced system efficiency.

Based on the Taylor expansion, the increase in $F(t)$, $\Delta F = F(t) - F(t-1)$, is represented as

$$\begin{aligned} \Delta F &= \sum_{k=1}^K \frac{\partial F(t-1)}{\partial T(k)} [T(k; t) - T(k; t-1)] + O\left(\frac{1}{t_c^2}\right) \\ &= \frac{1}{t_c} \sum_{k=1}^K \sum_{f=1}^F \sum_{b=1}^B \frac{\rho_k}{T^\alpha(k; t-1)} R_{f,b}(k|U_{f,b}, P_{f,b}; t) \\ &\quad - \frac{1}{t_c} \sum_{k=1}^K \rho_k T^{1-\alpha}(k; t-1) + O\left(\frac{1}{t_c^2}\right). \end{aligned} \quad (25)$$

For maximizing ΔF , we can neglect the second term, which is not a function of $U_{f,b}$ and $P_{f,b}$, and the third term $O(1/t_c^2)$ when $t_c \gg 1$. Furthermore, when we assume a per beam/frequency block power constraint, $U_{f,b}$ and $P_{f,b}$ can be decided independently for each beam of the respective frequency blocks. Therefore, the resource allocation policy for beam b of frequency block f selects user set $U_{f,b}(t)$ and allocation power set $P_{f,b}(t)$ at time t according to the following criterion.

$$G_{f,b}(U, P; t) = \sum_{k=1}^K \frac{\rho_k}{T^\alpha(k; t-1)} R_{f,b}(k|U, P; t). \quad (26)$$

$$\{U_{f,b}^*(t), P_{f,b}^*(t)\} = \arg \max_{\{U,P\}} G_{f,b}(U, P; t). \quad (27)$$

Term $G_{f,b}(U, P; t)$ is the resource allocation metric for user set U and power set P , and the combination of U and P that maximizes the resource allocation metric is selected.

In the uplink, transmission power $p_{k,f}$ can be independently determined for all users. In LTE and LTE-Advanced, $p_{k,f}$ for all users is predetermined apart from the user scheduling by the fractional power control [42], [43]. In such a case, the scheduling user set, $U_f^*(t)$, is selected to maximize $G_f(U, P; t)$ with a given $P_f(t)$.

In the downlink, the total transmission power of the base station is shared by the non-orthogonally multiplexed users. Therefore, U and P should be jointly optimized in principle. The resource allocation metric, $G_{f,b}(U, P; t)$, in (26) can be seen as the weighted sum of instantaneous user throughput $R_{f,b}(k|U, P; t)$ where the weighting factor for user k is $\rho_k/T^\alpha(k; t-1)$. Therefore, for given candidate scheduling user set U , the metric can be maximized by power allocation P , which maximizes the weighted sum of the instantaneous user throughput. We can use the iterative water-filling power allocation algorithm [24], [44] that achieves the maximum weighted sum of the user throughput by utilizing the uplink-downlink duality presented, for example, in [9] and [10]. With the optimum P for every given U , the maximization in (27) is performed with regard to U .

This kind of power allocation in the downlink based on the maximization of the weighted sum of the instantaneous user throughput has an additional merit when the channel coding block is distributed over multiple frequency blocks, as in LTE and LTE-Advanced [2]–[4]. In the downlink, the user order of decoding in the SIC should be in the order of the increasing normalized channel gain. This order may change among frequency blocks due to frequency-selective fading. In such a case, when the channel coding block is distributed over multiple frequency blocks, the decoding cannot be performed. This problem is solved by the fact that a user is allocated power only if all users with larger weighting factors have smaller channel gains. Since the weighting factor, $\rho_k/T^\alpha(k; t-1)$, is common to all frequency blocks, the decoding order of users, to which the power is allocated, becomes the same for all frequency blocks. Actually, the decoding order of users is equivalent to the order of the decreasing weighting factor, $\rho_k/T^\alpha(k; t-1)$, [10].

4. Numerical Results

4.1 Simulation Assumptions

This section shows the system-level throughput gain of the proposed NOMA scheme compared to OMA. Table 1 gives the simulation parameters.

These parameters basically follow the evaluation assumptions in 3GPP LTE [45]. However, we use the Shannon capacity-based throughput calculation and assume all cells have a single sector in this paper. The resource al-

Table 1 Simulation parameters.

Cell layout		Hexagonal 19-cell model
Inter-site distance		0.5 km
Overall transmission bandwidth		4.32 MHz
Frequency block bandwidth, W		180 kHz
Number of frequency blocks		24
Number of user terminals per cell, K		Parameterized from 5 to 40
Base station antenna	Number of antennas, M	2 for uplink evaluation 1 and 2 for downlink evaluation
	Antenna gain	14 dBi
User terminal antenna	Number of antennas, N	1 for uplink evaluation 2 for downlink evaluation
	Antenna gain	0 dBi
Base station transmission power		40 dBm
Maximum transmission power of user terminal		24 dBm
Distance-dependent path loss		$128.1 + 37.6 \log_{10}(r)$ dB, r : kilometers
Lognormal shadowing		Standard deviation = 8 dB Correlation among cells = 0.5
Instantaneous fading		Six-path Rayleigh, rms delay spread = 1 μ s, f_D = 55.5 Hz
Receiver noise density		-169 dBm/Hz
Scheduling interval		1 ms
Throughput calculation		Based on Shannon formula (Max. 6 b/s/Hz)
Averaging interval of user throughput		100 ms (t_e = 100)

Fig. 9 Cumulative distribution of downlink user throughput.

location described in Sect. 3.3, in which ρ_k for all users is assumed to be one and α is assumed to be a variable parameter, is conducted every millisecond. In the downlink, any kind of $M \times B$ -dimensional beamforming matrix determination criteria can be applied to the proposed NOMA scheme using the intra-beam superposition coding and SIC described in the previous section. In the following evaluations, we employ open loop-based random beamforming [46], [47]. Random beamforming is effective in reducing the channel state information feedback. In the uplink, fractional power control [42] assumed in LTE and LTE-Advanced is employed. In the following evaluations, the maximum number of non-orthogonally multiplexed users per beam of each frequency block, $N_{\max}$, is parameterized. The $N_{\max}$ of one corresponds to the OMA based on OFDMA as in LTE and LTE-Advanced.

4.2 Simulation Results

Figure 9 shows the cumulative distributions of the downlink user throughput. The $N_{\max}$ values of one, two, and four are evaluated. The values of M and B are set to two. Term K is 30 and α is set to 1.0.

Fig. 10 Downlink total user throughput as a function of K .

Figure 9 clearly shows that the proposed NOMA ($N_{\max} > 1$) scheme achieves better throughput than OMA for the entire region of the cumulative distribution. This is because the user throughput of OMA is severely limited by the orthogonal bandwidth allocation, which reduces the bandwidth for the respective users. NOMA allows for wider bandwidth usage of all users irrespective of their channel conditions. Allocating high power to the power-limited cell-edge users associated with the SIC process, which is applied to the bandwidth-limited cell-interior users, enhances the throughput of the users under a wide range of channel conditions. The impact of the transmission bandwidth limitation on OMA is especially clear in the high cumulative distribution probability region, where the users are under bandwidth-limited conditions. Therefore, since we assume that the maximum throughput per Hertz is limited to 6 b/s/Hz (corresponding to 64QAM) in the simulation, the user throughput with OMA is limited to approximately 2 Mb/s. The gain by further increasing $N_{\max}$ from two to four is relatively small. This indicates that it is sufficient to multiplex non-orthogonally a moderate number of users to obtain the most from the potential gain of NOMA. It should be noted that the overhead required for the transmission of a downlink reference signal for the proposed NOMA scheme using intra-beam superposition coding and the SIC is the same as that for OMA irrespective of the $N_{\max}$ value. In the following, the user throughput value at the cumulative probability of 5% is denoted as the cell-edge user throughput [45].

Figure 10 shows the downlink total user throughput as a function of the number of users per cell, K . We tested the cases with $N_{\max}$ of one and two and evaluated $M = B$ of one and two. Assuming the same $M (= B)$, the proposed NOMA scheme using the $N_{\max}$ of two significantly increases the total user throughput compared to OMA. Interestingly, the total user throughput of NOMA with $M=1$ is very similar to that of OMA with $M=2$. This implies that NOMA using power-domain multiplexing has an effect similar to spatial multiplexing and achieves a level of performance that is competitive with that for OMA while using

Fig. 11 User throughput gain in downlink.

a smaller number of base station transmitter antennas. It should also be noted that as $M (= B)$ and $N_{\max}$ increases, the required number of K for obtaining a sufficiently saturated total user throughput is increased. This is because as M and $N_{\max}$ increase, we need more candidate users in order to select an appropriate user set to be scheduled to achieve a sufficient multiuser diversity gain.

Figure 11 shows the user throughput gain of downlink NOMA ($N_{\max} = 2$) relative to OMA at the respective user coverage positions. The value of the user coverage position corresponds to the cumulative probability of the user throughput. Therefore, roughly said, the user coverage position of 0 indicates the cell boundary and that of 1 indicates the vicinity of the base station. The values of M and B are set to two. Term K is set to 30. In OMA, α is assumed to be 1.0. For the proposed NOMA scheme, α of 1.0 and 1.5 are tested. The total user throughput values of OMA with $\alpha = 1.0$ and NOMA with $\alpha = 1.0$ and 1.5 are approximately 31, 47, and 40 Mb/s, respectively. Figure 11 shows that a user throughput gain of greater than one is achieved for the entire region of the user coverage position by using the proposed NOMA scheme. We see that the cell-edge user throughput gain is especially large, which means significant improvement in the user fairness. When α is 1.0, the proposed NOMA scheme achieves approximately 1.5 and 1.2-fold gains in the total user throughput and cell-edge user throughput, respectively, compared to OMA. When α is increased to 1.5 for the proposed NOMA scheme, the proposed scheme increases the throughput gain at the cell-edge to approximately 1.6-fold at the cost of a moderate reduction in the total user throughput, which however is still 1.3-fold higher than that of OMA with $\alpha = 1.0$. This is because when α is increased, more radio resources tend to be allocated to the cell-edge users.

Figure 12 shows the uplink cell-edge and average user throughput levels as a function of $N_{\max}$. The number of users per cell, K , is 30. The weighting factor, α , is set to 1.0. NOMA employing the MMSE-SIC simultaneously improves the cell-edge and average user throughput compared to OMA ($N_{\max} = 1$), similar to the downlink case. This is because NOMA employing the MMSE-SIC allows

Fig. 12 Uplink cell-edge and average user throughput as a function of $N_{\max}$.

Fig. 13 User throughput gain in uplink.

for wider bandwidth usage for all users irrespective of their channel conditions, and the MMSE-SIC process enhances the throughput of users over a wide range of channel conditions. As $N_{\max}$ is increased, both the throughput measures are improved, however the additional gains tend to be saturated. When $N_{\max}$ is 4, NOMA with the MMSE-SIC increases both the cell-edge and average user throughput by approximately 1.4 and 1.7-fold, respectively, compared to OMA.

Figure 13 shows the user throughput gain of uplink NOMA with the MMSE-SIC relative to OMA at multiple user coverage positions. For OMA, α is set to 1.0. For NOMA employing the MMSE-SIC, $N_{\max}$ is set to 4 and α of 1.0 and 2.0 are tested. NOMA with the MMSE-SIC achieves a user throughput gain higher than one regardless of the user position. NOMA employing the MMSE-SIC with $\alpha = 2.0$ improves the user fairness within a cell compared to the case of $\alpha = 1.0$, but at the cost of a moderately reduced average user throughput. When $\alpha = 2.0$, NOMA employing the MMSE-SIC increases the cell-edge and average user throughput at the same time by approximately 1.6 and 1.4-fold, respectively, compared to OMA.

5. Conclusion

Aiming at further enhancement of the system efficiency and QoE, especially at the cell edge, in future radio access, this paper presented our investigation on NOMA. We proposed a NOMA scheme that uses intra-beam superposition coding and the SIC, for the cellular MIMO downlink. The basic resource allocation (bandwidth and power) problem for NOMA was also discussed. Simulations indicated that NOMA with a moderate number of non-orthogonally multiplexed users significantly enhances the system-level throughput performance in both the down- and uplink compared to OMA, which is widely used in 3.9 and 4G mobile communication systems. In this paper, we focused on the merits of NOMA in terms of the achievable throughput gains. The other potential merits of NOMA for practical system design are listed in [27], [31]. Based on these results, NOMA with an advanced receiver such as the SIC can be viewed as a promising candidate multiple access scheme for future radio access.

To further verify the effectiveness of NOMA, performance evaluations that assume practical channel coding and data modulations, and the impact of residual interference in the SIC process are essential. In [27], [30]–[32], performance evaluations that assume practical channel coding and data modulation are shown. In [32], [33], [35], the impact of the residual interference due to channel estimation and decoding error in the SIC process on the achievable throughput performance is investigated. In [24], [31], more practical resource allocation methods (compared to the method in Sect. 3.3) are investigated. In the downlink, the SIC process at the user terminal requires modulation and coding scheme information of not only its own signal but also other user signals, which are superposition coded. A realistic physical/media access control (MAC)-layer control signaling method that allows for the SIC process at the user terminal should be investigated. In the uplink, since the SIC process is implemented at the base station, we do not see a significant increase in the signaling overhead by introducing NOMA. In addition, the conventional control signaling assumed in LTE or LTE-Advanced may be reused in a straightforward manner. However, the designs of the uplink reference signal for channel estimation and uplink power control should be investigated in order to accommodate multiple user transmissions within the same frequency block. These issues are left for future study.

References

- [1] Y. Kishiyama, A. Benjebbour, H. Ishii, and T. Nakamura, "Evolution concept and candidate technologies for future steps of LTE-A," Proc. IEEE ICCS2012, pp.473–477, Singapore, Nov. 2012.
- [2] 3GPP TS36.300, Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description.
- [3] 3GPP TR36.913 (V8.0.0), "3GPP; TSG RAN; Requirements for further advancements for E-UTRA (LTE-Advanced)," June 2008.

- [4] 3GPP TR36.814 (V9.0.0), "Further advancements for E-UTRA physical layer aspects," March 2010.
- [5] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*, Cambridge University Press, 2005.
- [6] P. Bergman, "Random coding theorem for broadcast channels with degraded components," *IEEE Trans. Inf. Theory*, vol. IT-19, no. 3, pp. 197–207, March 1973.
- [7] M. Costa, "Writing on dirty paper," *IEEE Trans. Inf. Theory*, vol. 29, no. 3, pp. 439–441, May 1983.
- [8] G. Caire and S. Shamai (Shitz), "On the achievable throughput of a multi-antenna Gaussian broadcast channel," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1692–1706, July 2003.
- [9] P. Viswanath and D.N.C. Tse, "Sum capacity of the vector Gaussian broadcast channel and uplink-downlink duality," *IEEE Trans. Inf. Theory*, vol. 49, no. 8, pp. 1912–1921, Aug. 2003.
- [10] N. Jindal, S. Vishwanath, and A. Goldsmith, "On the duality of Gaussian multiple-access and broadcast channels," *IEEE Trans. Inf. Theory*, vol. 50, no. 5, pp. 768–783, May 2004.
- [11] W. Yu and J.M. Cioffi, "Sum capacity of Gaussian vector broadcast channels," *IEEE Trans. Inf. Theory*, vol. 50, no. 9, pp. 1875–1892, Sept. 2004.
- [12] H. Weingarten, Y. Steinberg, and S. Shamai (Shitz), "The capacity region of the Gaussian multiple-input multiple-output broadcast channel," *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 3936–3964, Sept. 2006.
- [13] B. Hassibi and M. Sharif, "Fundamental limits in MIMO broadcast channels," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 7, pp. 1333–1344, Sept. 2007.
- [14] A.D. Wyner, "Shannon-theoretic approach to a Gaussian cellular multiple-access channel," *IEEE Trans. Inf. Theory*, vol. 40, no. 6, pp. 1713–1727, Nov. 1994.
- [15] O. Somekh and S. Shamai, "Shannon-theoretic approach to a Gaussian cellular multiple-access channel with fading," *IEEE Trans. Inf. Theory*, vol. 46, no. 4, pp. 1401–1425, July 2000.
- [16] B. Farhang-Boroujeny, "OFDM versus filter bank multicarrier," *IEEE Signal Process. Mag.*, vol. 28, pp. 92–112, May 2011.
- [17] V. Vakilian, T. Wild, F. Schaich, S. ten Brink, and J.F. Frigon, "Universal filtered multi-carrier technique for wireless systems beyond LTE," *Proc. IEEE Globecom Broadband Wireless Access Workshop*, pp. 223–228, Atlanta, U.S.A., Dec. 2013.
- [18] G. Wunder, P. Jung, M. Kasparick, T. Wild, F. Schaich, Y. Chen, S. ten Brink, I. Gaspar, N. Michailow, A. Festag, L. Mendes, N. Cassiau, D. Ktésau, M. Dryjanski, S. Pietrzyk, B. Eged, P. Vago, and F. Wiedmann, "5G NOW: Non-orthogonal, asynchronous waveforms for future mobile applications," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 97–105, Feb. 2014.
- [19] F. Rusek and J.B. Anderson, "Constrained capacities for faster-than-Nyquist signaling," *IEEE Trans. Inf. Theory*, vol. 55, no. 2, pp. 764–775, Feb. 2009.
- [20] L. Ping, L. Liu, K. Wu, and W.K. Leung, "Interleave division multiple-access," *IEEE Trans. Wireless Commun.*, vol. 5, no. 4, pp. 938–947, April 2006.
- [21] L. Ping, Q. Guo, and J. Tong, "The OFDM-IDMA approach to wireless communication systems," *IEEE Trans. Wireless Commun.*, vol. 14, no. 3, pp. 18–24, June 2007.
- [22] T. Matsumoto, Y. Hatakeyama, and S. Konishi, "Performance analysis of interleave-division multiple access for uplink in multi-cell environment," *Proc. IEEE IWCNC2011*, pp. 376–381, Istanbul, Turkey, July 2011.
- [23] S. Tomida and K. Higuchi, "Non-orthogonal access with SIC in cellular downlink for user fairness enhancement," *Proc. IEEE IS-PACS2011*, pp. 1–6, Chiangmai, Thailand, Dec. 2011.
- [24] N. Otao, Y. Kishiyama, and K. Higuchi, "Performance of non-orthogonal access with SIC in cellular downlink using proportional fair-based resource allocation," *Proc. IEEE ISWCS2012*, pp. 476–480, Paris, France, Aug. 2012.
- [25] Y. Endo, Y. Kishiyama, and K. Higuchi, "Uplink non-orthogonal access with MMSE-SIC in the presence of inter-cell interference," *Proc. IEEE ISWCS2012*, pp. 261–265, Paris, France, Aug. 2012.
- [26] J. Umehara, Y. Kishiyama, and K. Higuchi, "Enhancing user fairness in non-orthogonal access with successive interference cancellation for cellular downlink," *Proc. IEEE ICCS2012*, pp. 324–328, Singapore, Nov. 2012.
- [27] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for future radio access," *Proc. IEEE VTC2013-Spring*, pp. 1–5, Dresden, Germany, June 2013.
- [28] K. Higuchi and Y. Kishiyama, "Non-orthogonal access with random beamforming and intra-beam SIC for cellular MIMO downlink," *Proc. IEEE VTC2013-Fall*, pp. 1–5, Las Vegas, U.S.A., Sept. 2013.
- [29] Y. Hayashi, Y. Kishiyama, and K. Higuchi, "Investigations on power allocation among beams in non-orthogonal access with random beamforming and intra-beam SIC for cellular MIMO downlink," *Proc. IEEE VTC2013-Fall*, pp. 1–5, Las Vegas, U.S.A., Sept. 2013.
- [30] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System-level performance evaluation of downlink non-orthogonal multiple access (NOMA)," *Proc. IEEE PIMRC2013*, pp. 611–615, London, U.K., Sept. 2013.
- [31] A. Benjebbour, Y. Saito, Y. Kishiyama, A. Li, A. Harada, and T. Nakamura, "Concept and practical considerations of non-orthogonal multiple access (NOMA) for future radio access," *Proc. IEEE IS-PACS2013*, pp. 770–774, Naha, Japan, Nov. 2013.
- [32] A. Benjebbour, A. Li, Y. Saito, Y. Kishiyama, A. Harada, and T. Nakamura, "System-level performance of downlink NOMA for future LTE enhancements," *Proc. IEEE Globecom2013*, pp. 66–70, Atlanta, U.S.A., Dec. 2013.
- [33] K. Yamamoto, Y. Saito, and K. Higuchi, "System-level throughput of non-orthogonal access with SIC in cellular downlink when channel estimation error exists," *Proc. IEEE VTC2014-Spring*, pp. 1–5, Seoul, Korea, May 2014.
- [34] H. Katayama and K. Higuchi, "Non-orthogonal access with SIC using inter-cell interference coordination based on coordinated power control for cellular uplink," *Proc. IEEE VTC2014-Spring*, pp. 1–5, Seoul, Korea, May 2014.
- [35] N. Nonaka, A. Benjebbour, and K. Higuchi, "System-level throughput of NOMA using intra-beam superposition coding and SIC in MIMO downlink when channel estimation error exists," *Proc. IEEE ICCS2014*, pp. 1–5, Macau, Nov. 2014.
- [36] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control," *IEEE/ACM Trans. Netw.*, vol. 8, no. 5, pp. 556–567, Oct. 2000.
- [37] A. Sang, X. Wang, M. Madihian, and R.D. Gitlin, "A flexible downlink scheduling scheme in cellular packet data systems," *IEEE Trans. Wireless Commun.*, vol. 5, no. 3, pp. 568–577, March 2006.
- [38] F. Kelly, "Charging and rate control for elastic traffic," *European Transactions on Telecommunications*, vol. 8, pp. 33–37, 1997.
- [39] A. Jalali, R. Padovani, and R. Pankaj, "Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system," *Proc. IEEE VTC2000-Spring*, pp. 1854–1858, Tokyo, Japan, May 2000.
- [40] R. Knopp and P.A. Humblet, "Information capacity and power control in single cell multiuser communications," *Proc. IEEE ICC'95*, pp. 331–335, Seattle, U.S.A., June 1995.
- [41] D. Bertsekas and R. Gallager, *Data Networks*, Prentice-Hall, 1992.
- [42] W. Xiao, R. Ratasuk, A. Ghosh, R. Love, Y. Sun, and R. Nory, "Uplink power control, interference coordination and resource allocation for 3GPP E-UTRA," *Proc. IEEE VTC2006-Fall*, pp. 1–5, Montréal, Canada, Sept. 2006.
- [43] 3GPP, TS 36.213, "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures," May 2009.
- [44] M. Kobayashi and G. Caire, "An iterative water-filling algorithm for maximum weighted sum-rate of Gaussian MIMO-BC," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1640–1646, Aug. 2006.

- [45] 3GPP, TR 25.814 (V7.0.0), "Physical layer aspects for Evolved UTRA," June 2006.
- [46] P. Viswanath, D.N.C. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Trans. Inf. Theory*, vol.48, no.6, pp.1277–1294, June 2002.
- [47] M. Sharif and B. Hassibi, "On the capacity of MIMO broadcast channels with partial side information," *IEEE Trans. Inf. Theory*, vol.51, no.2, pp.506–522, Feb. 2005.

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.). In 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 Tokyo University of Science. He is currently an Associate Professor at Tokyo University of Science. 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, a recipient of the Young Researcher's Award from the IEICE in 2003, the 5th YRP Award in 2007, and the Prime Minister Invention Prize in 2010. He is a member of the IEICE and IEEE.

Anass Benjebbour received his B.E., M.E., and Dr. Eng. degrees in communications and computer engineering in 1999, 2001, and 2004, respectively, all from Kyoto University, Japan. Since 2004 he has been with NTT DOCOMO Inc. R&D Japan, where he has been actively involved in the development of system concepts and the design of technologies for future radio access. He was also involved in 3GPP standardization for LTE Releases 10 and 11. He received the best paper award at IEEE PIMRC in 2002

and the young researcher's award from IEICE in 2006. He is a senior member of IEEE.