Invited Review Paper

Overview of nonlinear signal processing in 5G and 6G access technologies

Eiji Okamoto


1 Department of Electrical and Mechanical Engineering, Graduate School of Engineering, Nagoya Institute of Technology
Gokiso-cho, Showa-ku, Nagoya-shi 466-8555, Japan

a) okamoto@nitech.ac.jp

Received March 2, 2021; Revised April 12, 2021; Published July 1, 2021

Abstract: In recent years, there have been several advances in wireless communication systems, including the increase in transmission rate, and as a result, they are playing an increasingly significant role in our lives. To meet the widespread application scope, the commercialization of the fifth-generation mobile communications system (5G) has been initialized. This paper gives an overview of the 5G system and discusses the nonlinear signal processing technologies that support its performance improvement. Here, the term “nonlinear signal processing” is defined as an algorithm for transmitting and receiving data in communications that uses a nonlinear mechanism. Because there is a significant shortage of frequency bands in wireless communications, advanced technologies are being integrated to maximize frequency utilization. In addition, unlike in the 4G era, there are additional use cases that require performance guarantees for quality, delay, and the number of multiple connections in the 5G system, and the performance requirements for wireless systems are becoming more stringent. Nonlinear signal processing is key to improving the performance of these systems, and this paper outlines the nonlinear technology used in 5G wireless access. We will introduce multiple-input multiple-output transmission and non-orthogonal multiple access (NOMA), which will help improve the performance of wireless systems. In addition, we introduce a radio wave-encrypted NOMA with a physical layer security that is proposed by the authors. Furthermore, the technical trend of 6G is briefly introduced.

Key Words: 5G, nonlinear signal processing, massive machine type communication, multiple-input multiple-output, non-orthogonal multiple access

1. Introduction

In 2020, the 5th generation mobile communications system (5G) was launched in Japan, and 5G terminals are gradually spreading with the expansion of the service area. The fact that 5G is faster than the fourth-generation mobile communications systems (4G) outlines the evolution of the previous generation networks, but unlike the previous evolution, this network can accommodate communication with users other than mobile terminals such as automobiles and the Internet of things (IoT) terminals [1, 2]. 5G will provide a massive platform for remote work and education, which has almost
become a necessity across the globe, amid the spread of the coronavirus. Thus, the early spread and deployment of 5G is desired. In this paper, we will explain several aspects of 5G, which is the first mobile communication system to add the use cases of performance guarantee for quality, latency, and number of connections due to the demand for advanced applications. This feature is different from the use case of 4G and is a highly challenging service in terms of the principles of wireless communication. Given that its realization requires the application of advanced technologies, nonlinear signal processing was adopted. In this paper, we focus on the nonlinear signal processing that supports 5G technology and introduce the motivation for its application along with a few ways to improve its performance. Here, the term “nonlinear signal processing” is defined as an algorithm for transmitting and receiving data in communications that uses a nonlinear mechanism. In the past, the capability of signal processing was limited in wireless communication systems, especially in user terminals, and then, linear algorithms were used to eliminate signal interference among users and to reduce the decoding complexity in the receiver. However, by using nonlinear signal processing, the channel capacity of communication can be increased and the decoding error rate can be reduced with the same frequency bandwidth. In recent years, because the frequency band has become extremely tight, nonlinear signal processing has been actively used in 5G. We specifically outline the multiple-input multiple-output (MIMO) transmission [3, 4] and non-orthogonal multiple access (NOMA) [5, 6], which are key to improving the performance of these systems. Numerical simulation results show that the capacity and transmission quality can be improved by enhancing signal processing. In addition, we introduce a radio wave encryption NOMA with physical layer security [7, 8] and verify that it achieves both encryption and an increase in the number of connected users.

We introduce the elements that outline the performance of mobile communication systems in Section 2 and describe the new use cases and technical issues in 5G in Section 3. The results of the integration of nonlinear signal processing are described in Section 4, and the specific signal processing in MIMO transmission and NOMA is introduced in Section 5. The 6G trends are briefly discussed in Section 6, and the conclusion is drawn in Section 7.

2. Key performance indicator of mobile communication systems

Wireless transmission is convenient because it has features that wired transmission does not have, such as portability, which allows the communication terminal to be carried around, and broadcastability, which allows information to be sent over a wide area at the same time. Wireless systems are widely used today, not only in cell phones and smartphones but also in satellite broadcasting.

Figure 1 shows the number of users subscribed to mobile phone systems in Japan. The number of cell phone users has constantly increased since the launch of the first-generation mobile communication system (1G) (car phone) service in the 1980s due to its convenience. At present, the number of mobile terminal subscriptions in Japan is about 1.7 per capita, and the subscription rate is still rising. The number of subscriptions also includes terminal-to-terminal communication such as vending machines equipped with communication devices, and given the arrival of the IoT era, it is likely that this number will continue to increase in the future.

![Fig. 1. Number of subscribers to cell phone systems in Japan.](image)
Table I. Transmission rate and spectral efficiency of third and fourth generation mobile communication systems.

| Generation | Release 6 HSPA (3.5G) | Release 8 LTE (3.9G) | LTE-Advanced (4G) |
|------------|------------------------|----------------------|-------------------|
| Peak rate  | DL 14.4 Mbps           | 326.4 Mbps           | 1 Gbps            |
|            | UL 5.7 Mbps            | 86.6 Mbps            | 500 Mbps          |
| Peak spectral efficiency bit/s/Hz | DL 3                  | 15                   | 30                |
|            | UL 2                  | 3.75                 | 15                |

The theoretical error-free maximum transmission speed of transmission signals in communications is given by

\[ C = B \log_2(1 + S/N) \text{ bit/s,} \]  

(1)

where \( C \) is the channel capacity, \( B \) Hz is the bandwidth of the signal, and \( S/N \) is the signal to noise ratio (SNR) of the receiver [9]. Thus, to increase the transmission speed of communication, it is necessary to increase the bandwidth of the signal or the SNR of the receiver. However, since the transmission medium of wireless communication is space, unlike wired communication using copper wires or optical fibers, the attenuation of the received power with respect to the propagation distance is large, and the power of the received signal is significantly degraded. Generally, in plane wave propagation in a vacuum, the power decay is proportional to the square of the distance. For example, when a plane wave propagates between antennas that are separated by a distance of 100 m, the received power is attenuated by a factor of 10,000, and the received SNR is degraded.

On the other hand, according to Eq. (1), if the bandwidth \( B \) of the signal is increased, the transmission speed can be increased in proportion to \( B \). However, since the transmitted waves are radio waves, if two transmitters send out radio waves of the same frequency in the same space and at the same time, the received signal is a combination of both waves. It is impossible to separate the two waves, and the receiver will not be able to correctly retrieve the information from either signal. This is called interference. Therefore, it is necessary to use different frequencies for different purposes. The Ministry of Public Management, Home Affairs, Posts and Telecommunications (MIC), Japan, defines the use and bandwidth of radio waves in the Radio Law [10], and radio waves below 10 GHz are almost fully associated for various applications. In other words, radio waves are a precious resource, and it is not easy to increase the size of \( B \). Therefore, the transmission speed of wireless communication is generally less than 1/1000 of that of wired communication.

Therefore, an important performance improvement index for wireless communication systems is the frequency efficiency (or spectral efficiency) bit/sec/Hz, that is, the improvement of the transmission rate per 1 Hz. Table I shows the peak transmission rate and frequency efficiency for the 3.5–4th generation as specified by the 3rd Generation Partnership Project (3GPP) [11], a standardization project that is studying 5G standards [12, 13]. In the table, the DL and UL indicate the communication direction of the downlink from the base station to the terminal, and vice versa for the uplink. As seen in the table, the peak rate of the system increases dramatically as the generation progresses, but at the same time, the transmission rate per 1 Hz also increases significantly. In other words, improving frequency efficiency is key for the evolution of wireless communication systems. With regard to the aforementioned factors, to meet the demand for the constantly increasing number of terminals and to improve the transmission rate, the mobile communication system goes through a major update approximately every 10 years. The second generation was released in the early 1990s, and the third and fourth generations were released in the 2000s and early 2010s, respectively, with the fifth generation being released in 2020 [14–16]. Because the update of the generation is carried out in line with the development of the technology at the given time, the mechanism of communication often changes drastically. However, to prevent interruptions in the services provided, frequent generation releases are scheduled, which ensure backward compatibility and smooth transitions between any two generations. Based on this scheduling, the study of the sixth-generation mobile communication system (6G) has already started across the world including Japan.
3. 5G challenges

For all communications systems up to 4G, only one scenario was assumed for generational evolution, which was the increase in the speed of cell phones and smartphones. However, for 5G, two new scenarios have been added in addition to the existing one. Figure 2 shows the three scenarios and their primary applications in the IMT Vision Recommendation [17] issued by the International Telecommunication Union Radiocommunication Sector (ITU-R). Moreover for communications systems up to 4G, only the enhanced mobile broadband (eMBB) scenario, shown in the upper part of the figure, was envisioned. In 5G, the eMBB scenario is also being developed, and the peak transmission rate is increased to 20 Gbps, which is significantly greater than that of 4G (1 Gbps). As shown in the figure, this will enable three-dimensional video, 4K/8K high-definition video distribution, virtual reality, and augmented reality applications.

For 5G, the ultra-reliable and low-latency communications (URLLC) and massive machine-type communications (mMTC) scenarios shown in the lower part of the figure have been newly added. In the URLLC scenario shown in the lower right of the figure, the performance requirement is aimed at enabling the transmission of 32-byte packet data with a success probability of 99.999% or higher within a wireless interval of 1 ms or less, and this scenario realizes highly reliable and low-latency transmission [18]. In fact, achieving excellent quality is a strict requirement for wireless communications, and as a result, 5G is undergoing a major transformation. The reasons for this are explained below.

Figure 3 shows a schematic diagram of radio wave propagation. Because radio wave propagation exists over a wide range, radio waves from a base station to a terminal are not only received as
direct waves but also as reflected, diffracted, scattered, and reflected mountain waves. Given that the
direction of reception, arrival time, and Doppler effect of these radio waves are different, the combined
radio waves at the terminal may strengthen or weaken each other, and the received power changes at
every instance depending on the location, time, and frequency. This phenomenon is called multipath
fading, and the effect of fading sometimes causes reception errors in wireless communications. This is
the reason why only eMBB-equivalent scenarios were provided until 4G, and therefore URLLC requires
the application of advanced technologies to realize highly reliable and low-latency communications
with excellent quality while considering the effects of fading. However, the use cases for this technology
are expected to expand, and the applications and services that can be realized are expected to grow
significantly. The future applications will include communication for autonomous driving and remote-
controlled driving, control of non-wired devices in smart factories, robot control, and telemedicine.

On the other hand, the mMTC shown in the lower left of Fig. 2 is a super-multiple-connected
scenario that must accommodate one million connections per km². If a large number of terminals
communicate using the same frequency at the same time, interference will occur. Therefore, to ensure
that multiple terminals are connected simultaneously, it is necessary to shift the radio resources
slightly (described later in Section 5.3). The mMTC is realized by applying this technology, and this
allows an extremely large number of terminals to connect simultaneously. As shown in the figure, the
main use case is for information collection and control of IoT terminals distributed over a wide area
in a smart city and home appliances in a smart home, and relatively low-speed transmission of small
amounts of data to and from a large number of terminals is assumed.

The addition of the URLLC and mMTC scenarios is expected to contribute to the development of
various industries. The 5G service, which started in March 2020, has already provided eMBB and
some URLLC services, and full-scale URLLC and mMTC services are expected to be available around
2023. Therefore, 5G will not be fully involved in new industries for several years to come, and we will
gradually be able to witness several 5G innovations in the forthcoming years.

4. Improving performance of 5G multi-user access with nonlinear
signal processing

In wireless communication systems, even though various technologies are integrated to improve effi-
ciency, due to limitations in computing power and power consumption, signal processing designs are
primarily based on linear calculations with relatively low complexity. However, due to the continuous
improvement in the performance of semiconductors, the signal processing capability of the receiving
side has been dramatically improved. As a result, nonlinear signal processing has been actively used
in wireless access network technology.

Table II shows the classification of linear and nonlinear signal processing in wireless access network
technology. The axes of resources used in the design are carrier wave (in-phase (I), quadrature (Q)),
time, frequency, space, and spreading code. The advantage of the orthogonal norm is the avoidance of
interference due to the individual distribution of the allocated resources and the resulting reduction
in the amount of decoding computation. For example, as shown in Fig. 4, the quadrature phase-shift
keying (QPSK) is an orthogonal superposition of binary phase-shift keying (BPSK) on the I- and Q-
axes of the carrier wave, and decoding can be performed independently on both axes for simplification.
In contrast, the main benefit of non-orthogonal norms is that the capacity of the communication
channel can be expanded by increasing the signal density more than the orthogonal norm. If we

| Specification  | Signal processing | Pros                                      | Cons                                |
|----------------|-------------------|-------------------------------------------|-------------------------------------|
| Orthogonal     | Linear            | Reduced decoding complexity and interference | Limitations of quality and capacity  |
| Non-orthogonal | Nonlinear         | Capacity expansion and higher quality through diversity | Increased decoding complexity and interference |
Fig. 4. Examples of orthogonal and non-orthogonal assignment in carrier axis; (a) orthogonal assignment, (b) non-orthogonal assignment.

Fig. 5. Optimal allocation of wireless resources in heterogeneous networks.

consider the example of the carrier axis in Fig. 4, it corresponds to the case where Gaussian mapping, which is an integrated signal point arrangement of the IQ axis and is performed by combining the bit strings transmitted by the I and Q axes. In other words, the diversity effect is achieved because the transmitted signals are widely distributed (with interference) on the resource axis. This simply outlines the fact that a Gaussian distribution is optimal for input signals that satisfy Shannon’s channel capacity [19]. However, to achieve this benefit, integrated maximum likelihood decoding that includes interfering signals is necessary at the receiver side, similar to Shannon’s theory. In the previous example, the received points corresponding to the correlated bitstrings must be collectively estimated as a maximum likelihood sequence estimation, otherwise, the bit error rate will be worse than that of QPSK. Because the amount of calculation increases exponentially with the number of received points that are to be decoded together, it is much larger than that of QPSK, where the I- and Q-axis can be determined for each received point. Because of the complexity and volume of computation required on the decoding side, nonlinear signal processing methods have not been emphasized in conventional wireless communication systems. However, due to changes such as the tightening of frequencies and the improvement of semiconductor performance, it is gradually being used more actively.

As shown in Fig. 5 as an example, in 4G and latter communication systems, multiple base stations are superimposed and heterogeneous networks are used to communicate with user equipment (UE) using the same frequency. Because multiple small cells are superimposed on a macrocell with a large coverage area, there is a non-orthogonal resource partition in the spatial and frequency axes. The users in the vicinity of a small cell are connected to a small cell base station (BS) to obtain good reception power and can communicate at high speed. However, when the transmission power of the BS increases, the SNR of the accommodating UE improves, but the interference power of neighboring UEs connected to other BSs also increases. Thus, the signal to interference plus noise ratio (SINR) of the
neighboring UEs decrease and the communication quality deteriorates. In addition, the superposition of super wide area cells with satellites, unmanned aerial vehicles, or drones, etc. as base stations, shown in the figure, will be considered for future applications. In this case, if the objective function is to maximize the capacity of the entire system under the constraint of constant transmitting power, this becomes a nonlinear combinatorial optimization problem in radio resources and transmit power allocation of each BS or UE. This mainly depends on the channel coefficient between all BSs and UEs.

Because the number of combinations is very high, a sub-optimal solution is considered for practical applications. In the future, as described in Section 6, a large-capacity optimal multisource connection technology with low computational complexity will be required not only for a group of BSs in a narrow area but also for multiple BSs in a wide area. In Beyond 5G, adaptive radio resource allocation on the carrier, time, frequency, space, and code axes, and the allocated power axis, is required to improve the system efficiency.

The following are some examples of 5G wireless access technologies.

5. Signal processing for 5G user access technology

5.1 MIMO multiplexing transmission

Figure 6 shows a transmission method wherein the number of antennas of the transmitter and receiver are multiplexed to send data, and this method is called MIMO multiplexing transmission [3, 4]. In this case, if the number of antennas on the transmitter and receiver sides is $N_t$ and $N_r$, respectively, the communication channel capacity $C$ can be increased by a factor of $\min(N_t, N_r)$ without the need to increase the signal bandwidth. This is a significant advantage in the current situation where the radio is used for various purposes and frequencies are being exhausted. This has also become a standard feature in mobile communication systems since 3.9 G, and is a mandatory technology for 4G and beyond. As described in the next section, the same effect can be obtained even if there are $N_t$ or $N_r$ terminals with one antenna either on the transmitting or receiving side. By extending this principle, 5G uses massive MIMO transmission, wherein the number of antennas at the base station is increased to tens or hundreds, and simultaneous connection with a large number of receivers, or high-speed communication to a small number of receivers equipped with many antennas is achieved.

When the complex channel coefficients $h_{ij}$ between each transmitting and receiving antenna in Fig. 6 are

$$H = \begin{pmatrix} h_{11} & \cdots & h_{1N_r} \\ \vdots & \ddots & \vdots \\ h_{N_t1} & \cdots & h_{N_tN_r} \end{pmatrix},$$

the relationship between the transmitter and receiver signals can be expressed as $y = Hx + n$. Here, $x$ and $y$ are the transmitting and receiving signal vectors, respectively, and $n$ is the received noise vector. In this method, transmission signals are superimposed non-orthogonally on the spatial axis. However, because the signals of all transmitting antennas except the $i$-th antenna interfere with the $i$-th receiving antenna due to non-orthogonal multiplexing, it is necessary to remove the interference by signal processing at the receive side. In this case, if $N_r \geq N_t$, the degree of freedom of the receiving antenna is equal to or greater than that of the transmitting antenna, the signals of each transmitting antenna can be orthogonalized by spatial filtering. First, we orthogonalize the received signal by linear filtering with $\hat{x} = \mathbf{w}_r y$. The simplest case is a matched filtering (MF), wherein $\mathbf{w}_r = H^H$ is used as the receiver weight. Here, $H$ is the Hermitian transpose. The case where the receiver

![Fig. 6. Transmitter and receiver design for MIMO multiplexing transmission.](image)
weight is $w_r = H^H (HH^H + \delta I_{N_r})^{-1}$ is called zero-forcing (ZF) filtering, wherein $\delta$ is a constant, $I_{N_r}$ is an $N_r$-dimensional identity matrix, and $\sigma^2_N$ is the variance of the noise at the receiver. The following equation shows the result of the minimum mean square error (MMSE) equalization when $w_r = H^H (HH^H + N_r \sigma^2_N I_{N_r})^{-1}$. The obtained $\hat{x}$ is orthogonalized or quasi-orthogonalized between the antennas. Due to the orthogonalization, each element of $\hat{x}$ can be demapped independently, and we can decode the signal with the $N_tQ$ number of searches for $Q$-ary modulation. The performance of the filtering is $MF < ZF < MMSE$, which is contrasting to the computation involved.

Next, let us consider decoding by nonlinear signal processing. Here, we use the joint full-search of

$$\hat{x} = \arg\min_x ||y - Hx||^2$$

as the maximum likelihood decoding (MLD), which is an optimal decoding method for reducing the received bit error rate. However, the amount of exploration for decoding is $Q^{N_t}$, which increases exponentially with the number of transmitting antennas.

Figure 7 shows the bit error rate (BER) characteristics at the receiver side in a Rayleigh fading channel when $N_t = N_r = 2$. It is seen that the nonlinear signal processing of MLD has better characteristics than the linear signal processing of ZF and MMSE equalization. This is because of the integrated decoding of the signal spread on the spatial axis. As a result, in MIMO multiplex transmission, the capacity of the communication channel can be increased by transmitting non-orthogonally on the spatial axis, and the quality can be improved by compensating the increase in calculation complexity in the MLD, which is the nonlinear signal processing, on the receiver side.

5.2 Multiple access using massive MIMO

In the coming years, given the manner of advancements and exhaustive researches, it is likely that a system will be constructed such that information from a large number of IoT sensor devices distributed in the city will be transmitted to a cloud server and accumulated as big data. Further, artificial intelligence (AI) will be used to analyze and extract features from this data using machine learning and other techniques, making it possible to obtain minute and detailed information pertaining to the trends, features, and concepts that could not be studied previously. By feeding this information back to the system devices and IoT terminals to control physical systems in the real world, an advanced information processing cycle will be realized. This system combines real and cyberspace and is called the cyber-physical system (CPS). This system is sometimes referred to as a digital twin system for similar concepts. By utilizing the CPS cycle, an intelligent information and communication fusion space with an extremely high time and space resolution can be developed, and it is thought that it will be possible to realize unprecedented services and applications. This will realize the safety, security, and comfort of society in an ultra-smart society.
In general, the accuracy of AI is proportional to the volume of data used, and to improve the accuracy, it is necessary to collect sensing data from IoT terminals over a wider area. However, the power supply for IoT terminals is not always sufficient. For example, for illuminance and temperature sensors in the field, the frequency of information transmission can be low, but they need to operate on a battery-powered source for 10 years. The mMTC scenario described in Section 3 suffices these requirements. Figure 8 shows a conceptual diagram of the connection of a large number of IoT terminals in the uplink from the terminal to the BS. Several MIMO principles are used to expand the number of antennas in the base station to several hundreds, which further enables a simultaneous connection of up to several hundred single-antenna IoT terminals without the need to increase the required frequency. Furthermore, by increasing the number of connections with the NOMA method described in the next section, and by switching the IoT terminals to be connected by time division, the number of IoT terminals that can be accommodated by a single base station can be increased to tens of thousands. In addition, the mMTC provides a standard that reduces the transmission speed by decreasing the frequency width, which in turn enables power-saving communication.

However, when the number of users in the uplink is large, the $N_t$ in Fig. 6 increases to several hundreds, thus resulting in the problem of increased computational requirements at the receiver. If linear filtering is used, the order of the inverse matrix operation in Eq. (2) increases significantly. Furthermore, if the MLD of Eq. (3) is used, the computational complexity increases to an extent that detection cannot be achieved in real-time. Therefore, various quasi-maximum likelihood detection methods suitable for massive MIMO using linear or nonlinear signal processing have been proposed [20]. For example, in the case of linear filtering, a method to calculate the approximate inverse matrix of $H$ and reduce the computational complexity has been proposed. In addition, many methods for quasi-maximum likelihood sequence estimation similar to MLD have been proposed. Ones of these methods are

- A typical method is to start with an initial solution by linear filtering and searching for more accurate solutions in the neighborhood through descent, random, or tabu search.

- By using the principle of belief propagation, the log likelihood ratio (LLR) is recursively exchanged between the received signal and the candidate transmitted signal. Subsequently, quasi-maximum likelihood sequence estimation is obtained.

- When the number of active transmitted signals is smaller than the number of antennas, $H$ can be regarded as sparse, and $\hat{x}$ is estimated by using compressive sensing technology.

- Machine learning is applied for MIMO detection and $H$ estimation [21–23] to reduce the computational complexity.

5.3 Non-orthogonal multiple access
The orthogonal frequency division multiple access (OFDMA) method, which uses the principle of orthogonal frequency division multiplexing (OFDM) for multiple access, has been adopted in the fourth-generation system (3.9 generation in the official standard), and the service was launched in
Figure 9 illustrates the principle of the OFDM method, which assigns slightly different frequencies (channels) to each user, but uses the OFDM principle to synchronize each channel. This eliminates the need for guard bands, which are usually inserted between channels to prevent interference, and results in a tightly packed orthogonal arrangement of channels. This is optimal in terms of the orthogonal arrangement, and it also improves the frequency efficiency. In addition, because the system is a packet switching system and not a call switching system with fixed users and channels, the capacity of the entire system can be increased by adaptively assigning each channel to a user with good propagation conditions between the terminal and the BS. This is referred to as multiuser diversity. By combining the multiple access technology with the MIMO transmission technology described above, a transmission rate of 100 Mbps or higher in the downlink has been achieved in 4G.

After dividing the radio resources in a similar way, if orthogonal division is performed for each user, there will be no mutual interference, and since no calculation is required to extract the allocated resources, the calculations on the decoding side can be reduced. In the early stages of 4G, such a resource allocation was necessary because of the limited computing power of UEs and BSs. However, with the increase in the number of subscribers and the demand for high capacity in mobile communication systems, orthogonal resource partitioning is insufficient for radio resources even for tight-lattice partitioning such as OFDMA.

To achieve higher transmission rates, the non-orthogonal multiple access (NOMA) was adopted in 5G systems [5, 6] as shown in Fig. 10. As seen in the figure, multiple users are assigned to the power axis of the same channel. For example, two users with different propagation conditions, such as near and far from the BS, are selected as a pair, and the transmitting BS increases the power of the signal for the farther user and superimposes it on the near user signal for transmission. A user in the vicinity of the BS, which is the receiving side, receives the mixed signals of the neighboring and distant users, wherein the signal of the distant user can be received with high power, and thus, this signal is first determined correctly. Further, by applying successive interference cancellation (SIC), the superimposed interference can be accurately removed, and only the signal addressed to the neighboring user is extracted. On the other hand, the distant user also receives the mixed signals of the neighboring and distant users, but due to the large attenuation of the propagation path, the interference of the neighboring user small to an extent that it can be ignored, and only the signal addressed to the distant user, which is assigned high power, is extracted. This method establishes a multiway connection that accommodates more users and has higher efficiency than Fig. 9. The principle of this method contributes to the speed-up of eMBB and the realization of multiple connections of mMTC.

As a concrete example of the operating principle of NOMA, we explain the downlink transmission in a 5G cellular system. Figure 11 shows a conceptual diagram of the downlink NOMA system. In the downlink from the BS to the UE, the BS is first informed of the propagation path information (channel state information) of the user in the target cell by feedback. Based on this information, the BS allocates each subband (subcarriers) to the users while considering the proportional fairness (PF). This operation is called frequency scheduling. In this process, the NOMA allocates the frequency bands while allowing overlapping users that may cause interference. The average user capacity when
all subbands are allocated is maximized according to the PF standard. In the example shown in the figure, UE1,4 and UE3,2 are superimposed in subbands 1 and 3, respectively, near and far from the BS. By setting an appropriate difference in the transmission power, the capacity of the system can be increased compared to OFDMA.

Figure 12 shows a comparison of the normalized channel capacity of UE1 and 4 in NOMA and OFDMA, where the propagation of conditions of UE1 is 100 times than those of UE4, i.e., 100 times higher channel power [6]. As seen, the sum of the channel capacities of both users is increased by using NOMA compared to OFDMA. Because the bandwidth used is the same, the frequency efficiency is considerably improved.

5.4 Spreading-based non-orthogonal multiple access for mMTC

In mMTC, wireless connections to a large number of IoT terminals are envisioned [1, 17, 24]. To achieve this, a method using NOMA technology as described above has been proposed [25–29]. In NOMA, a single radio resource is shared by multiple users, which causes mutual interference. However, overload transmission can be realized, wherein more users transmit than the number of radio resources, thus improving the performance of mMTC. Currently, orthogonal superposed connection-based methods such as enhanced MTC (eMTC), which was developed for 4G, is used for IoT connectivity, but 5G mMTC is more suitable for IoT connectivity, especially in the high-SNR regions on the receiver side. It is known that the capacity of the 5G mMTC system is larger than that of existing orthogonal systems in that region [30]. Various NOMA methods for mMTC have been proposed, depending on the resource axis they are superimposed on [28]. A few examples of this method are the low-density signature (LDS) [25] and sparse code multiple access (SCMA) [26], which spread the signal on the time or frequency axis. By sparsifying the transmitted samples, it is possible to superimpose a large number of users—for example, the case with four samples and six users enables 150% overload transmission with the number of users compared to the number of transmitted samples. In addition, by spreading the radio resources, the mutual interference due to non-orthogonal superposition can be suppressed, which enables a highly accurate user detection.

Figure 13 shows the block diagram of the mMTC NOMA transmission method in the uplink link wherein information is transmitted from the largest $J$ users to the base station. Each user maps
the transmitted bits using methods such as LDS and SCMA, and performs OFDM transmission by applying inverse fast Fourier transform (IFFT) and placing the mapped signals on the same subcarriers.

Each user encodes the transmitted bit-sequence with an outer concatenation code, divides it into \( \log_2 M \) bits along the modulation multiplicity \( M \), applies NOMA mapping, and transmits it simultaneously. We denote the transmitted bits of user \( j \) by \( b_j = (b_{1j}, \ldots, b_{(\log_2 M)j}) \), \( b_{lj} \in \{0, 1\} \), where \( l = 1, \ldots, \log_2 M \). Let \( x_j = (x_{1j}, \ldots, x_{Kj})^T \) be the codebook corresponding to \( b_j \), where \( K \) is the number of transmitted samples. Then, the transmitted bit sequence is converted into a complex signal of \( N_j \) dimensions, where \( N_j < K \), with a multidimensional signal point mapping \( c_j = g_j(b_j) \) of

\[
c_j = g_j(b_j)
\]

and

\[
c_j = (c_{1j}, \ldots, c_{N_j})
\]

is obtained. Subsequently, by using a spreading matrix with null insertion, which is referred to as the binary mapping matrix \( V_j \), we obtain \( x_j \) as \( x_j = V_j c_j = V_j g_j(b_j) \). For the LDS, the existing digital modulation mapping \( g_j \) is used for \( c_j \), e.g., for \( M = 4 \), as in Fig. 4(a)

\[
c_{ij} \in \left\{ \pm \frac{1}{\sqrt{2}}, \pm \frac{j}{\sqrt{2}} \right\}
\]

is used. Therefore, \( x_j \) corresponds to the transmission sequence of the QPSK signal code-spread by the binary mapping matrix \( V_j \). A generalized suboptimal codebook generation method for SCMA is described in detail in [31]. Although \( K \geq J \) was required up to 4G, transmission with \( K < J \) is achieved with the NOMA. To illustrate this, we show the mapping for the case where the resources used by the user are guaranteed in advance, which is referred to as the “grant type”. When the factor graph matrix \( F \) given by

\[
F = (f_1, \ldots, f_J), \quad f_j = \text{diag}(V_j V_j^T)
\]

is used, a 150% overloaded transmission with a grant type \( K = 4 \) samples and \( J = 6 \) users is achieved with the 4 x 6 matrix \( F \) of

\[
F = \begin{pmatrix}
0 & 1 & 1 & 0 & 1 & 0 \\
1 & 0 & 1 & 0 & 0 & 1 \\
0 & 1 & 0 & 1 & 0 & 1 \\
1 & 0 & 0 & 1 & 1 & 0
\end{pmatrix}
\]

In \( F \), the column \( j \) represents the non-zero transmit sample position of user \( j \), and the column weight of \( F \) is \( N_j \). The row \( k \) represents the superimposed user in the \( k \)-th transmission sample. Thus, in

Fig. 13. Block diagram of uplink SCMA, LDS, and SCCMA transmissions.
of \( F \) of (8), each user sends \( N_j = 2 \) non-zero samples out of \( K = 4 \) samples, and three users are always superimposed on each transmission sample. As a result, \( K \) samples are transmitted for each of the \( M \)-ary transmitted bit strings, and the codebook is sparse in that only \( N_j \) points are non-zero. In addition, the sample position to be transmitted for each user is determined by \( V_j \), and the sparse sample point transmission reduces the number of users that cause overlapping interference on the receiving side.

On the other hand, in the “grant-free” transmission method, the \( N_j \) and non-zero sample positions are determined by the transmitting user. In this case, the maximum number of transmitting users \( J \) is not limited to six in (8), and user superposition can be easily performed. However, the number of superpositions in the transmitted sample increases with \( J \), which results in a tradeoff of increased users and the suppression of inter-user interference, which causes the user detection performance to deteriorate. In general, the grant-free transmission is ideal for transmission error rate performances for increasing \( N_j \) when the number of users \( J \) transmitting at the same time is small, and to reduce \( N_j \) to less than half of \( K \) to mitigate interference when \( J \) is large.

In Fig. 13, OFDM transmission is performed for \( x_{kj} \) samples of \( x_j \) placed on the same \( K \) subcarriers for all user \( j \). As a result, overload transmission is possible, and the transmission efficiency can be increased as compared to OFDMA. At the receiver side, after domain transformation is performed using fast Fourier transform (FFT), the multiuser detection method is used to collectively decode the user signals and calculate the LLR, which is then input to the outer channel code decoder at the later stage. The updated LLR result is then returned to the multiuser detector to perform turbo iterative decoding.

5.5 Chaos-based grant-free sparse chaos code multiple access scheme

We proposed a new method called grant-free sparse chaos code multiple access (GF-SCCMA) [7, 8], which is similar to the LDS and SCMA methods. The GF-SCCMA is a non-orthogonal multiple access method that can ensure security in the physical layer by generating a codebook using chaotic signals. This ensures only the pairs that share the chaotic initial value as a common key can decrypt. In addition, it does not require a detailed design of the codebook in advance, and a large number of codebooks can be created, thus enabling GF transmission for an arbitrary number of users. Figure 14 shows an example of the SCCMA codebooks [7]. The codebook is a Gaussian modulation, which is based on the Shannon theory as shown in Fig. 4(b), and the system capacity can be increased as shown in Fig. 16 below. In the receiver side of the uplink transmission, Gaussian belief propagation (GaBP) is applied for decoding, and the system capacity is increased such that \( K = 32 \) and \( J = 64 \).

In the following section, we show the numerical results of transmission performance. The existing methods include grant-free low-density signature (GF-LDS), wherein the sending user selects an arbitrary column of the grant-type factor graph matrix \( F \), GF-SCMA, OFDMA without the grant-free transmission, and grant-type LDS and SCMA using Eq. (8). Table III shows the basic simulation parameters. In this study, the number of antennas for both the UE and the BS was assumed to be one, and the performances were mainly evaluated for the number of transmission samples \( K = 4 \), number of active transmission users \( J_A = J = 6 \), number of subcarriers 2048, and the turbo code length.
Table III. Basic simulation parameters.

| Multiple access scheme | OFDMA          | (GF-)LDS, SCMA | GF-SCCMA       |
|-------------------------|----------------|---------------|---------------|
| Modulation              | QPSK           | QPSK (LDS),   | Chaos-based   |
|                         | $M = 4$        | Codebook (SCMA) | codebook $M = 4$ |
| No. of samples          | $K = 4$        |               |               |
| No. of maximum users    | $J = 6$        |               |               |
| No. of max weights in $F$| -              | $(J, 2)$      |               |
| Transmission            | Multi-carrier OFDM w/ round-robin scheduling |               |               |
| No. of subcarriers      | 2048           |               |               |
| No. of data subcarriers | 2048 = 4 sample x 512 symbol |               |               |
| No. of guard interval   | 8 samples      |               |               |
| Cell layout             | Non-sectorized hexagonal single cell model |               |               |
| User location           | Randomly distributed |               |               |
| Channel                 | Path loss exp. = 3.5, Standard deviation of shadowing loss = 7 dB, 1-dB decaying 8-path i.i.d. quasi-static Rayleigh fading |               |               |
| Channel estimation      | perfect        |               |               |
| Multi-user detection algorithm | Subcarrier extraction | Message passing algorithm (MPA) using sum product algorithm |               |
| No. of detection iteration | -              | 15 fixed      |               |
| Channel coding          | Rate 1/2 punctured turbo code, code length = 1024 |               |               |
| Channel decoding        | BCJR MAP decoding with 8 iteration |               |               |
| Error detection         | CRC-16         |               |               |
| Outer decoding iteration | -              | 1             |               |

Fig. 15. Comparison of normalized throughput performances versus cell-edge average received SNR in single cells in fading channel.

1024. The common key signals between the transmitter and receiver sides of GF-SCCMA method were assumed to be shared in advance. The number of non-zero samples for the SCCMA method was fixed at $N_j = 2$. In the grant-free transmission, the resource information used by the receiver, i.e., the detection of the column number of $F$, is assumed to be perfect. The turbo codewords are placed in the frequency direction in one OFDM frame. One turbo iteration in decoding was performed between the message passing algorithm (MPA)-based user signal detector and the turbo decoder. The data to be turbo coded is CRC-16 encoded in advance, and the normalized throughput characteristic is calculated using the detection result of the CRC-16 code, where the sample transmission per subcarrier is one. This indicates that the normalized throughput is defined as one for error-free orthogonal transmission.

Figure 15 shows the transmission characteristics in a single cell and fading environment. The
convergence of MPA is improved due to the instantaneous variation of received power caused by the user distribution and fading, and that the performance of the NOMA with the grant type is improved from the low SNR region, as compared to OFDMA. Although the performance of the proposed GF-SCCMA method is degraded as compared to that of the grant method, which is caused by the degradation of MPA detection by grant-free, it is confirmed that the proposed GF-SCCMA method shows better performance than the GF-LDS and GF-SCMA methods where the codebooks completely overlap in the region where the average SNR is more than 5 dB. This is because the LLR output of the GF-SCMA method has a more desirable distribution for the subsequent turbo decoder, based on Shannon theory. In addition, since the saturation throughput of the GF-SCCMA method is 1.5, it is superior to OFDMA above a certain SNR. In addition, GF-LDS and GF-SCMA do not reach the saturation throughput when SNR < 20 dB.

The figure also shows the signal separation characteristics when the real and imaginary parts of the chaotic initial key signals of all users are shifted by $10^{-3}$ at the receiving base station. In this case, the throughput is 0. This indicates that the decoder without the key signal cannot separate the signals correctly and the physical layer security is ensured. Therefore, the proposed method can achieve both physical layer security and channel coding gain in the uplink environment, contrary to the existing GF-based and OFDMA methods.

Figure 16 shows the performance of the proposed method in a single-cell fading environment where the average SNR received at the cell edge is 16 dB, and the number of active transmitting users is $J_A$. In this case, all transmissions are GF transmissions. From the results, it is clear that the characteristics of the GF-SCCMA method are better than those of the existing GF-based methods for all the number of users, thus indicating that chaos-based NOMA has a significant throughput improvement effect. This is because the GF-SCCMA signal is closer to a Gaussian distribution as described in Section 4. Consequently, this example showed that the nonlinear signal processing was effective in 5G.

In this study, we assumed that the channel coefficients and the resource index for each user, i.e., the active user detection, are perfect at the receiver side. However, in reality, this needs to be done by control signal transmission before data transmission. In the mMTC, this active $J_A$ UE detection should be done from several thousand potential $J$ UEs with an active ratio with a small percentage, and their channels need to be estimated using a non-orthogonal preamble. This active user detection also incorporates the techniques of nonlinear optimization problems [32–34].

6. Prospects on Beyond 5G and 6G

Figure 17 illustrates the concept of cell-free operation in Beyond 5G and 6G. The evolution of 5G,
which is currently under consideration by 3GPP and other organizations, could be diversified by the use cases beyond eMBB, URLLC, and mMTC. To support the evolution of such peak elemental performances, it is necessary to increase the transmission rate of the uplink and use radio waves in the high-frequency band, including the THz band for a wider $B$. However, as the use of higher frequency bands increases, the range of radio waves at the same transmission power becomes relatively shorter, and thus, the existing cellular concept shown on the left of the figure will require a cell-free mechanism, as shown on the right, which provides flexible radio access in terms of time, space, and frequency. As a result, there is an increasing need for algorithms that perform areal optimization rather than single control such as optimization of resource allocation at several BSs. In addition, because the optimal solution is subject to combinatorial explosion, various methods are being investigated to reduce the amount of computation and obtain an effective solution using machine learning and other techniques [35].

7. Conclusions
In this paper, we presented an overview of 5G and outlined the nonlinear signal processing that supports the improvement of the performance of 5G. The 5G communication system is expected to open up newer industrial applications because it has use cases that guarantee the quality, latency, and performance of the number of connections, unlike 4G. The nonlinear signal processing is necessary to achieve the performance improvement, and the nonlinear technologies used for 5G wireless access were explained, including the MIMO transmission and NOMA. In addition, our proposal of NOMA with radio wave encryption was introduced. Since 6G is expected to implement nonlinear signal processing on a larger scale for optimization over a wider area, it is expected that wireless signal processing techniques for generating global semi-optimal solutions with low complexity, including the machine learning technique, will be further developed.

Acknowledgments
This work was partly supported by JSPS KAKENHI (grant number JP18K04136) and KDDI Foundation Research Grant Program. The author wishes to express his appreciation for the supports received.

References
[1] 5GMF White Paper, “5G mobile communications systems for 2020 and beyond,” [Online] http://5gmf.jp/wp/wp-content/uploads/2016/07/5GMF_WP101_All.pdf, ver. 1.0.1, July 2016.
[2] M. Shafi et al., “5G: A tutorial overview of standards, Trials, Challenges, Deployment, and Practice,” IEEE Journal on Selected Areas in Communications, vol. 35, no. 6, pp. 1201–1221, June 2017.
[3] G.J. Foschini, “Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas,” Bell Labs Technical Journal, vol. 1, no. 2, pp. 41–59, Autumn 1996.
[4] A. Goldsmith, S.A. Jafar, N. Jindal, and S. Vishwanath, “Capacity limits of MIMO channels,” IEEE Journal on Selected Areas in Communications, vol. 21, no. 5, pp. 684–702, June 2003.
[5] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, “Non-orthogonal multiple access (NOMA) for cellular future radio access,” 2013 IEEE 77th Vehicular Technology Conference (VTC Spring), pp. 1–5, June 2013.

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

[7] E. Okamoto, N. Horiike, and T. Yamamoto, “Sparse chaos code multiple access scheme achieving larger capacity and physical layer security,” Proc. International Symposium on Wireless Personal Multimedia Communications, pp. 1–7, December 2017.

[8] E. Okamoto, N. Horiike, and T. Yamamoto, “Large-scale grant-free sparse chaos code multiple access scheme for 5G IoT,” Proc. IEEE Vehicular Technology Conference 2018 Fall, pp. 1–5, August 2018.

[9] C.E. Shannon, “A mathematical theory of communication,” Bell Syst. Tech. J., vol. 27, pp. 379–423 and 623–656, 1948.

[10] [Online] https://www.tele.soumu.go.jp/e/index.htm

[11] [Online] https://www.3gpp.org/

[12] 3GPP TS 36.300, “Evolved universal terrestrial radio access (E-UTRA) and evolved universal terrestrial radio access network (E-UTRAN): Overall description: Stage 2,” April 2017.

[13] 3GPP, TR 36.814 (V9.0.0), “Further Advancements for E-UTRA Physical Layer Aspects,” March 2010.

[14] 3GPP TR 21.915, V15.0.0, “Release 15 description; Summary of Rel-15 work items (Release 15),” September 2019.

[15] 3GPP TS 23.501 version 15.3.0 Release 15, “System architecture for the 5G system,” September 2018.

[16] 3GPP TS 38.300 version 15.5.0 Release 15, “5G; NR; Overall description; Stage-2,” March 2019.

[17] ITU-R, “IMT Vision – Framework and overall objectives of the future development of IMT for 2020 and beyond,” Recommendation M.2083-0, September 2015.

[18] 3GPP TR 38.913, “Study on scenarios and requirements for next generation access technologies (Release 14),” V14.1.0, December 2016.

[19] C.E. Shannon, “A mathematical theory of communication,” The Bell System Technical Journal, vol. 27, no. 3, pp. 379–423, July 1948.

[20] M.A. Albreem, M. Juntti, and S. Shahabuddin, “Massive MIMO detection techniques: A survey,” IEEE Communications Surveys & Tutorials, vol. 21, no. 4, pp. 3109–3132, Fourthquarter 2019.

[21] N. Samuel, T. Diskin, and A. Wiesel, “Deep MIMO detection,” 2017 IEEE 18th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), pp. 1–5, 2017.

[22] H. He, C. Wen, S. Jin, and G.Y. Li, “A model-driven deep learning network for MIMO detection,” 2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP), Anaheim, CA, USA, pp. 584–588, 2018.

[23] Y. Huang, P.P. Liang, Q. Zhang, and Y. Liang, “A machine learning approach to MIMO communications,” 2018 IEEE International Conference on Communications (ICC), 2018.

[24] 3GPP TR22.861 V14.1.0, “FS_SMARTER - massive Internet of Things,” 2016.

[25] R. Hoshyar, F.P. Wathan, and R. Tafazolli, “Novel low-density signature for synchronous CDMA systems over AWGN channel,” IEEE Trans. Signal Processing, vol. 56, no. 4, pp. 1616–1626, April 2008.

[26] H. Nikopour and H. Baligh, “Sparse code multiple access,” Proc. IEEE Int’l Symp. Personal Indoor and Mobile Radio Communications (PIMRC), pp. 332–336, September 2013.

[27] C. Yan, Z. Yuan, W. Li, and Y. Yuan, “Non-orthogonal multiple access schemes for 5G,” ZTE Communications, vol. 14, no. 4, pp. 11–16, October 2016.

[28] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen, and L. Hanzo, “A survey of non-orthogonal multiple access for 5G,” IEEE Communications Surveys & Tutorials, vol. 20, no. 3, pp. 2294–2323, thirdquarter 2018.
[29] N. Horiike, E. Okamoto, and T. Yamamoto, “A downlink non-orthogonal multiple access scheme having physical layer security,” *EURASIP Journal on Wireless Communications and Networking*, vol. 2018, no. 1, pp. 1–11, August 2018.

[30] E. Okamoto, M. Mikami, and H. Yoshino, “Performance comparison of multi-user shared multiple access scheme in uplink channels,” *IEICE Trans. Commun.*, vol. E102-B, no. 8, pp. 1458–1466, August 2019.

[31] M. Taherzadeh, H. Nikopour, A. Bayesteh, and H. Baligh, “SCMA codebook design,” Proc. IEEE Vehicular Tech. Conf. (VTC-2014Fall), pp. 1–5, September 2014.

[32] L. Liu and W. Yu, “Massive connectivity with massive MIMO—Part I: Device activity detection and channel estimation,” *IEEE Transactions on Signal Processing*, vol. 66, no. 11, pp. 2933–2946, June 2018.

[33] T. Hara and K. Ishibashi, “Grant-free non-orthogonal multiple access with multiple-antenna base station and its efficient receiver design,” *IEEE Access*, vol. 7, pp. 175717–175726, November 2019.

[34] E. Okamoto, Y. Masuda, and T. Yamamoto, “Study on secure user detection and channel estimation scheme in grant-free massive machine type communications,” Proc. International Symposium on Wireless Personal Multimedia Communications (WPMC2020), pp. 1–2, October 2020.

[35] M.E. Morocho-Cayamcela, H. Lee, and W. Lim, “Machine learning for 5G/B5G mobile and wireless communications: Potential, Limitations, and future directions,” *IEEE Access*, vol. 7, pp. 137184–137206, 2019.