MIMO-based Grant-free Access: Enabling Diverse QoS in Massive and Low-latency Communication

Ameha T. Abebe & Chung G. Kang
School of Electrical Engineering, Korea University
Seoul, Korea, Republic of
Email:{ameha_tsegaye,ccgkang}@korea.ac.kr

***Abstract*—Grant-free (GF) access enables massive and low-latency access for various use cases of the 5th generation (5G) mobile communication and beyond. However, these use cases have diverse quality-of-service (QoS) requirements, which can be measured in terms of an access success rate and latency from a random access perspective. Consequently, a unifying GF scheme, which enables supporting of a diverse QoS, is highly desired. This paper proposes a GF access scheme in which high-QoS users superpose multiple preambles to achieve collision and multiple access interference diversity and ultimately improve their access success rate. We further show that in the presence of multiple antennas at the base station (BS) a low-complexity receiver correctly detects active preambles with a significantly high probability, even under severe multiple access interference caused by non-orthogonal preamble transmission. Simulation results demonstrate multiple orders improvement in terms of the access success rate for critical-QoS users, even under severe noise and multiple access contamination.**

Index Terms*—grant-free access, massive MIMO, ZadoffChu sequences, quality of service, uRLLC, mMTC

I. Introduction

In contrast to the four-step resource request and grant-based random access in the 3GPP's long-term evolution (LTE) standard, grant-free (GF) random access permits users to send the preamble and data signal in one shot. GF random access is considered as an enabler for ultrareliable and low-latency communication (uRLLC) and massive machine-type communication (mMTC), as it simultaneously reduces the access delay and signaling overhead, respectively [1-3]. The preamble transmission in GF random access serves to identify the existence of an active transmission in the course of active user detection (AUD), while estimating the channels of active users. Massive multiple-input multiple-output (mMIMO)-based GF access has gained considerable attention recently [1-3]. Preamble signals received by multiple antennas can be jointly processed to significantly improve the AUD performance. Moreover, by employing well-known techniques for receiver beamforming, superposed users' signals can be differentiated based on their channel vectors.

The quality of service (QoS) of a GF access scheme can be measured by reliability and latency metrics [4]. In this regard, we consider the GF access success rate (probability) as a key performance indicator (KPI) to measure QoS [3-4]. It is to be noted that the GF access success rate comprehensively measures the GF access scheme's capability of avoiding access collision while ensuring its accuracy in AUD, channel estimation, and data detection. In this respect, a unified GF scheme that guarantees different QoS (success rates) for the various services of 5G and B5G, including uRLLC, mMTC, and critical mMTC, would be significant. For example, a uRLLC service requires 99.999-99.99999% success rate while it is sufficient to provide only 99.9% for vast majority of mMTC use cases [4]. To provide ultrareliability in GF access, the transmission of multiple preamble sequences by a user is investigated in [3]. Therein, users transmit multiple preambles over multiple preamble transmission slots by randomly selecting them from a pool of orthogonal preambles. The use of multiple preambles provides collision diversity, i.e., it increases the probability of avoiding preamble collision in at least one of the preamble transmission slots and thereby improves the success rate of random access [3]. However, the preamble overhead is increased upon using multiple preamble slots. Furthermore, a GF system for users with varying QoS is yet to be investigated.

In this paper, we consider a mMIMO-based GF access where users transmit one or multiple preambles from a pool of non-orthogonal preamble sequences based on their respective QoS. In contrast to [3], in which multiple preambles from a user are transmitted in orthogonal consecutive time slots, a high-QoS user superposes its multiple preamble signals in the proposed scheme, thereby reducing signaling overhead. Many preamble sequences are generated from multiple-root ZadoffChu (ZC) sequences while allowing nonzero cross-correlation among preambles, i.e., using the non-orthogonal preambles. The use of ZC sequences as preamble sequences enables multipath (multi-tap) channel estimation and facilitates orthogonal-frequency division multiple access (OFDM)-based data transmission. Another significant limitation of the majority of the previous works on GF access is that they consider single-tap channel estimation [1-3] and their extension to multi-tap channel estimation is not straightforward. For example, [1-3] employed binary and Gaussian random sequences as preambles, which are not suitable for multipath channel estimation. The single-tap channel assumption may be reasonable for narrowband systems such as NB-IoT. However, for low-latency services, which require transmission on wider bandwidth and shorter time span, this assumption is not practical. ZC sequences, on the other hand, exhibit an excellent cross-correlation property that enables low-complexity multi-path channel estimation. Furthermore, a low-complexity AUD algorithm inspired by simultaneous-hard thresholding (SHT)

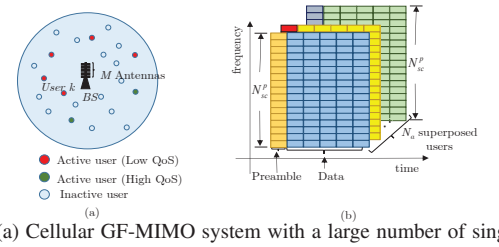


Fig. 1. (a) Cellular GF-MIMO system with a large number of single-antenna users and a BS with M antennas (b) Preamble and data transmission for GF access: Illustration of multiple UEs superposed for their transmissions

[5], which is suitable for delay-intolerant communication, is proposed. In particular, the preamble signal received by multiple antennas at a BS is jointly processed by modeling it as an multiple-measurement vector (MMV) class of compressive sensing problems. Furthermore, for the proposed AUD scheme, simulation results show that an exponential decrease in preamble misdetection rate as the number of antennas at the BS is increased. Moreover, it is shown that the GF access success rate of high-QoS users is significantly improved upon superposing and transmitting multiple preambles.

II. System Model & Problem Formulation

A. Preamble and Data Transmission

We consider a GF random access system, as illustrated in Fig. 1 (a), wherein there are N_{UE} users each with a single antenna and a BS equipped with M antennas. At a random access opportunity, suppose that N_a active users contend for random access, where $N_a \ll N_{UE}$, i.e., a small fraction of the total users are active. As illustrated in the figure, users can be classified as having one of two QoS levels, a high or low level. In general, more than two QoS levels can be considered. Fig. 1 (b) illustrates an OFDM system in which N_a active users are superposed on the same time and frequency resources. Therein, a time domain is divided into a preamble slot and data transmission slot (OFDM symbols). Furthermore, the number subcarriers for the preamble and data slots are given by N_{sc}^p and N_{sc}^d, respectively, in a frequency domain for an OFDM system.

An active user first selects a single or multiple preamble sequences from a pool of N_p preambles, which is a predefined set, $\{\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_{N_p}\}$, where the m-th preamble is an N_{ZC}-length ZC sequence denoted as $\mathbf{p}_m \in \mathbb{C}^{N_{ZC}\times 1}$ for $m = 1, 2, \cdots, N_p$. The preamble length N_{ZC} must be equal or close to the number of subcarriers, N_{sc}^p, in the preamble regime, as each element of the inverse discrete Fourier-transformed (IDFT) preambles is mapped to a resource element (single subcarrier). When N_{sc}^p is slightly greater than N_{ZC}, the lower subcarriers in the preamble slot can be repeated by copying the first $(N_{sc}^p - N_{ZC})$ elements of preambles. It is to be noted that a normalized ZC sequence have a unit auto-correlation and a zero and $1/\sqrt{N_{ZC}}$ cross-correlation values if they are generated from the same and different roots, respectively. Therefore, in order to average-out the MAI form other users' transmission, it is advantageous if a user chooses its multiple preambles from different roots.

Moreover, suppose that $\bar{\mathbf{h}}_k^{(t)} \in \mathbb{C}^{\tau\times 1}$ denotes a time-domain channel between the k-th user and t-th BS antenna. In particular, it is defined as $\bar{\mathbf{h}}_k^{(t)} = [\bar{h}_{k,0}^{(t)}, \bar{h}_{k,1}^{(t)}, \cdots, \bar{h}_{k,\tau-1}^{(t)}]^T$, wherein $\bar{h}_{k,\ell}^{(t)}$ represents a multipath path channel component for the ℓ-th tap with delay-spread length τ. Let $\mathcal{P}_k$ denote an index set of the preambles selected by the k-th user. Then, $p_k \triangleq |\mathcal{P}_k|$ corresponds to the number of preambles employed by the k-th user. The preamble signal received at the t-th antenna in the BS, after applying a discrete Fourier transformation (DFT), is given as

$$\mathbf{y}_p^{(t)} = \sum_{k=1}^{N_a} \sum_{m\in\mathcal{P}_k} \mathbf{p}_m \otimes \bar{\mathbf{h}}_k^{(t)} + \boldsymbol{\omega}_p^{(t)}, \; t = 1, 2, \cdots, M. \quad (1)$$

where $\otimes$ denotes a circular convolution operation. We assume that $\boldsymbol{\omega}_p \sim \mathcal{CN}(0, \sigma^2 \mathbf{I}_{N_{ZC}})$ in (1) is the ambient noise with power σ^2, where $\mathbf{I}_N$ denotes an $N_{ZC} \times N_{ZC}$ identity matrix. For the MIMO receiver, the preamble transmission in (1) can also be expressed in a matrix multiplication form as

$$\mathbf{Y}_p = \mathbf{PH} + \boldsymbol{\Omega}_p, \quad (2)$$

where $N_{ZC} \times M$ matrices, $\mathbf{Y}_p$ and $\boldsymbol{\Omega}_p$, are concatenations of the received signal and noise vectors in (1), i.e., $\mathbf{Y}_p = \left[\mathbf{y}_p^{(1)}, \mathbf{y}_p^{(2)}, \cdots, \mathbf{y}_p^{(M)}\right]$ and $\boldsymbol{\Omega}_p = \left[\boldsymbol{\omega}_p^{(1)}, \boldsymbol{\omega}_p^{(2)}, \cdots, \boldsymbol{\omega}_p^{(M)}\right]$, respectively. The preamble matrix $\mathbf{P} \in \mathbb{C}^{N_{ZC}\times \tau N_p}$ has N_p blocks of columns associated with each preamble sequence and their $(\tau - 1)$ cyclically-shifted versions. In particular, the preamble matrix $\mathbf{P}$ can be expressed as a concatenation of submatrices (blocks) as $\mathbf{P} = [\mathbf{P}_1, \mathbf{P}_2, \cdots, \mathbf{P}_{N_p}]$, where $\mathbf{P}_m = [\mathbf{p}_{m,0}, \mathbf{p}_{m,1}, \cdots, \mathbf{p}_{m,\tau-1}]$ is formed by setting the first column as $\mathbf{p}_{m,0} = \mathbf{p}_m$ and the other $(\tau - 1)$ columns set by circularly rotating $\mathbf{p}_m$ by one element at a time, i.e., $\mathbf{p}_{m,\ell} = circ(\mathbf{p}_m, \ell)$. Similarly, the channel matrix is defined as $\mathbf{H} \triangleq [\mathbf{h}^{(1)}, \mathbf{h}^{(2)}, \cdots, \mathbf{h}^{(M)}]$ where its t-th column $\mathbf{h}^{(t)}$ belongs to the channel vector associated with the t-th antenna. For the sake of generality, let $\mathcal{D}_m$ be a set of users that selected and transmitted the m-th preamble. The block-sparse vector $\mathbf{h}^{(t)} = [\left(\mathbf{h}_1^{(t)}\right)^T, \left(\mathbf{h}_2^{(t)}\right)^T, \cdots, \left(\mathbf{h}_{N_p}^{(t)}\right)^T]^T \in \mathbb{C}^{\tau N_p \times 1}$ holds the channel gain coefficients associated with each preamble, i.e., the m-th subsector (block of elements) denoted as $\mathbf{h}_m^{(t)}$ is set by $\sum_{k\in\mathcal{D}_m} \bar{\mathbf{h}}_k^{(t)}$ as superposition of the channels for users which selected the m-th preamble. If there is no collision while selecting m-th preamble, $|\mathcal{D}_m| = 1$. Note that each column of $\mathbf{H}$ is block-sparse in a sense that the number of nonzero values in it is less than its dimension, i.e., $\left\|\mathbf{h}^{(t)}\right\|_0 \le \tau N_a \ll \tau N_p$ for $t = 1, 2, \cdots, M$ and an ℓ_0-norm operator $\|\cdot\|_0$. Moreover, these nonzero values are located as a block of τ elements.

From the transmission model in (2), it is now obvious that the frequency-domain channel between user k and the t-th antenna along N_{sc}^d data subcarriers is given as $\hat{\mathbf{h}}_k^{(t)} = \mathbf{W}[(\mathbf{h}_k^{(t)})^T, \mathbf{0}]^T$, where $\mathbf{W}$ denotes an N_{sc}^d-point discrete Fourier transform (DFT) matrix whose element at n-th row and m-th element is given as $w_{m,n} = (1/N_{sc}^d) \exp\left(-j2\pi mn/N_{sc}^d\right)$.

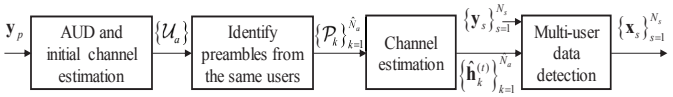


Fig. 2. Receiver structure for MIMO-based grant-free access in BS

Moreover, $[(\mathbf{h}_k^{(t)})^T, \mathbf{0}]^T$ is a $N_{sc}^d \times 1$ vector formed by zero-padding $\mathbf{h}_k^{(t)}$. Furthermore, the transmitted symbol vector is denoted as $\mathbf{x}_s \in \mathbb{C}^{N_a \times 1}$ with its k-th element holding the symbol transmitted by user k at the s-th subcarrier. Then the received data symbols vector $\mathbf{y}_s \in \mathbb{C}^{M \times 1}$ for subcarrier indices running as $s = 1, 2, \cdots, N_{sc}^d$, is given as

$$\mathbf{y}_s = \mathbf{H}_s \mathbf{x}_s + \boldsymbol{\omega}_s, \ s = 1, 2, \cdots, N_{sc}^d, \tag{3}$$

where $\boldsymbol{\omega}_s \sim \mathcal{CN}(0, \sigma_s^2 \mathbf{I}_M)$ is a noise vector associated with the s-th subcarrier of power σ_s^2.

III. Activity Detection and Channel Estimation

Given the above preamble and data transmission model, the receiver at the BS must first identify the active preambles (user transmissions), while estimating the corresponding channel, for data detection. Fig. 2 shows the MIMO-based receiver structure for the proposed GF access. It first performs AUD to identify the active preambles, which are then included in a set $\mathcal{U}_a$ for initial channel estimation. The receiver then attempts to identify preambles that belong to the same users by matching the estimated channel. Once the preambles belonging to the same users are identified, i.e., the sets $\{\hat{\mathcal{P}}_k\}$ for $k = 1, 2, \cdots, N_a$, then channel estimation can be performed. The second tier of channel estimation benefits from the multiple preamble consideration via averaging out of the possible error due to MAI and noise. Finally, on utilizing the channel estimates $\{\hat{\mathbf{h}}_k^{(t)}\}$, multiuser data can be detected via receive beamforming.

A. Active User Detection Based on SHT

In this study, we consider an AUD scheme inspired by the widely known algorithm referred to as simultaneous hard thresholding (SHT) or multichannel thresholding [5]. SHT is the simplest of the greedy algorithms, as it estimates the support of the signal in one shot as opposed to other greedy CS algorithms that involve iterative steps for support recovery. Given the preamble transmission model in (2), received preamble signals $\{\mathbf{y}_p^{(t)}\}_{t=1}^M$, and preamble matrix $\mathbf{P} = \left[\mathbf{P}_1, \mathbf{P}_2, \cdots, \mathbf{P}_{N_p}\right]$, the algorithm computes the correlation values and sums them along the M measurements. In particular, the correlation value corresponding to the m-th preamble is computed as

$$\xi_m = \sum_{t=1}^{M} \left\| \mathbf{P}_m^H \mathbf{y}_p^{(t)} \right\|_2^2, \ m = 1, 2, \cdots, N_p. \tag{4}$$

Then, the preambles with a correlation value greater than a predefined threshold ξ are included in the estimated support set $\mathcal{U}_a$. By employing a threshold value ξ, therefore, preambles with $\xi_m \geq \xi$ are declared as active. In the following sections, ξ is set to ensure that elements in the correct support set are also included in the estimated support, i.e., $\Lambda \in \mathcal{U}_a$. The initial channel estimation associated with the preambles in $m \in \mathcal{U}_a$, i.e., $\hat{\mathbf{h}}^{(t)} = [-\hat{\mathbf{h}}_m^{(t)} -]$ can be performed with least-square (LS)-based channel estimation as follows:

$$\hat{\mathbf{h}}^{(t)} = (\mathbf{P}_{\mathcal{U}_a})^\dagger \hat{\mathbf{y}}_p^{(t)}. \tag{5}$$

After the channel estimates associated with the detected preambles are computed, preamble indices, m and m' in $\mathcal{U}_a$, are considered to belong to the same users if their corresponding channel estimates, $\{\mathbf{h}_m^{(t)}\}$ and $\{\mathbf{h}_{m'}^{(t)}\}$, satisfy the following constraint:

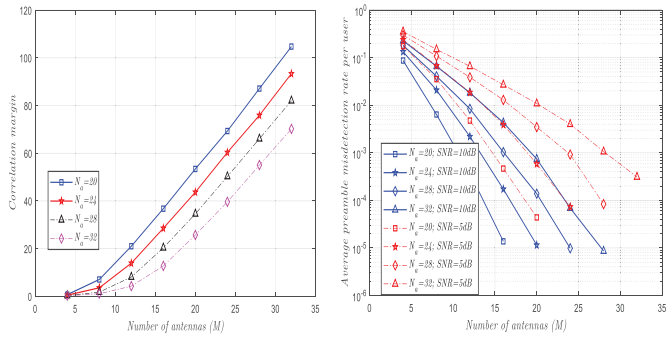
$$\left| \sum_{t=1}^{M} \left(\hat{\mathbf{h}}_m^{(t)} \right)^H \hat{\mathbf{h}}_{m'}^{(t)} \right| \geq \xi^{(2)}, \ \text{for } m, m' \in \mathcal{U}_a. \tag{6}$$

In (6), $\xi^{(2)}$ is considered as a threshold to declare whether two detected preambles belong to the same user. After grouping the detected preambles using (5), we obtain the estimates of the sets $\{\hat{\mathcal{P}}_k\}_{k=1}^{\hat{N}_a}$, where $\hat{N}_a$ is the number of detected active users. It should be noted that collided preamble transmissions are excluded in (6). For example, suppose $\mathcal{P}_k = \{m, m'\}$ and $\mathcal{P}_{k'} = \{m, m''\}$, implying that $\mathbf{p}_m$ is transmitted by two users, k and k'. Then, the channel estimates $\{\hat{\mathbf{h}}_m^{(t)}\}$ would differ considerably from $\{\hat{\mathbf{h}}_{m'}^{(t)}\}$ and $\{\hat{\mathbf{h}}_{m''}^{(t)}\}$. Thus, $\mathbf{p}_m$ would not be included in the estimated preamble sets, $\hat{\mathcal{P}}_k$ and $\hat{\mathcal{P}}_{k'}$.

IV. Simulation Results and Discussions

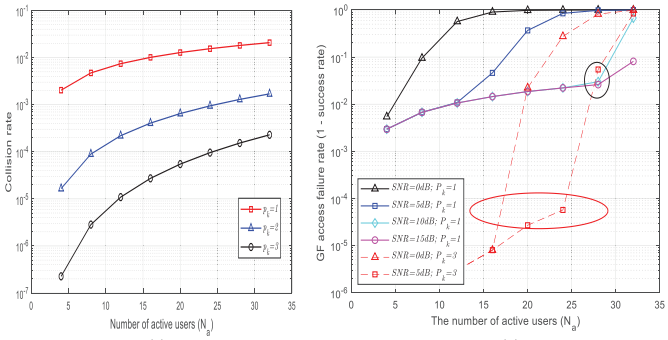
In this section, we discuss the performance of the proposed scheme with simulation results. For performance evaluation, we consider a GF random access resource (RAR) with an SC spacing of 15 kHz for both preamble and data transmission slots. The RAR constitutes $N_s = 128$ subcarriers in the frequency domain, with one preamble slot and two data OFDM symbols. In practice, the channel coherence time is much longer than that of two symbols, and the data transmission regime can carry more OFDM symbols. Furthermore, QPSK is employed for data transmission by all users, implying that the data regime carries 256 bits. A ZC sequence of $N_{ZC} = 127$ is mapped onto the 128 SCs in the preamble slot by circularly placing its first element to the last (128th) SC. A single BS with $M = 4 : 32$ antennas serves $N_a = 20 : 32$ active users. A total of $N_p = 1024$ unique preamble sequences are generated. Here, the maximum number of orthogonal (single-root) preamble sequences that can be generated is given as $\lfloor N_{ZC}/N_{SH} \rfloor = 42$ for $N_{ZC} = 127$ and $N_{SH} = \tau = 3$. Therefore, 1024 preambles are non-orthogonal with each other, as they are generated from $N_r = 25$ different roots. We consider a Gaussian channel between each user and each MIMO antenna with a flat power profile.

In Fig. 3 (a) and (b), the probability of misdetection (P_{mis}) and the average threshold margin (Δ_ξ) for variable number of users and receiver antennas (M) are presented. The threshold margin is defined as $\Delta_\xi \triangleq \min_{m \in \Lambda} \xi_m - \max_{m' \notin \Lambda} \xi_m$, i.e., the margin between the minimum and maximum correlation values, as computed in (4), of an active and inactive preamble have with the received signal, respectively. From the figure, it is



(a) (b)

Fig. 3. (a) average threshold margin vs. the number of active users (b) preamble misdetection rate vs. the number of antennas at the BS (M) $SNR = 5dB, 10dB$ and $N_a = 20 : 32$.



(a) (b)

Fig. 4. (a) collission rate vs. the number of active users (b) GF access failur rate vs. the number of users for mMTC and uRLLC users with $p_k = 1$ and $p_k = 3$ preambles per users, respectively.

observed that Δ_ξ increases linearly as M increases. For a Gaussian distributed noise and other-cell interference, it can be conjectured that a linear increase in Δ_ξ means an exponential decrease in the misdetection error. This conjecture is confirmed by Fig. 3 (b) wherein P_{mis} decreases exponentially (linearly in logarithmic scale) as the M increases. Therefore, it can be concluded that at the availability of a large number of antennas at the BS, a nearly perfect AUD can be ensured even under non-orthogonal preamble transmission.

Fig. 4 (a) presents the collision probability for a varying number of active users. The number of total preambles is maintained at $N_p = 1024$, and the number of active users is increased from 4 to 32. Furthermore, it is assumed that all users employ the same number of preambles, and the number of preambles per user assumes the values $p_k = 1, 2$, and 3. It is clear from Fig. 4 (a) that the collision probability is significantly reduced by multiple orders when multiple preambles are employed by each user. Fig. 4 (b) shows the GF access failure rate, $P_{failure}$, of the high- and low-QoS users to highlight the overall advantage of QoS differentiation in GF access. In this regard, we have implemented a 10-bit user ID embedded in the data transmission. Then, we assume that a user's packet is lost if this ID cannot be decoded. Both high- and low-QoS users employ a convolutional channel coding with a 1/2 code rate. The number of users is increased from $N_a = 4$ to $N_a = 32$, and the SNR ranges from 0 to 15 dB. High-QoS users constitute 20% of the users, and they randomly select, superpose, and transmit $p_k = 3$ preambles. In the figure, it is shown that when the SNR increases to 10 dB, the PDR decreases. Above SNR = 10 dB, however, the failure rate does not improve as other factors such as preamble collision and misdetection become detrimental factors, causing packet drop. In contrast, for high-QoS users, even at low SNR, i.e., 0 and 5 dB, a very low failure rate is observed. In fact, even if there exist 15 or 20 users contending for random access, a $P_{failure} = 10^{-4}$ can be experienced by high-QoS users, when the SNR is 0 or 5 dB, respectively. It is to be noted that this failure rate is within the failure rate that is recommend for uRLLC systems, i.e., 10^{-5}-10^{-3} [4]. The figure confirms that if multiple preambles are transmitted by a user with high QoS requirements, high reliability, i.e., low failure rate, may be achieved even despite severe noise and MAI.

V. Conclusion and Future Works

In this study, we presented a unified framework of grant-free access that can support users with diverse QoS requirements. The proposed GF access framework enables both mMTC and uRLLC use cases, particularly for their coexistence over a single radio resource, in the 5G or B5G mobile communication system. It is shown that the same framework supports a 99.9% to 99.99999% GF access success rate (reliability) depending on the QoS requirement of a user. Furthermore, it is observed that by permitting high QoS users to contend with superposition of multiple preambles, the grant-free access success rate can be improved by multiple orders. One important advantage of the proposed GF access framework is that it can be flexibly implemented in the current mobile communication standards, i.e., 3GPP LTE and NR specifications, with their moderate modification. In the future, we plan to investigate the optimization of the proposed scheme with respect to the various design parameters such as the number of preambles employed for high QoS users, preamble length, and the underlying overhead.

Acknowledgment

This work was supported in part by Samsung Research, Samsung Electronics and in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.2020R1A2C1009984).

References

[1] K. Senel and E. G. Larsson, "Grant-Free Massive MTC-Enabled Massive MIMO: A Compressive Sensing Approach," in *IEEE Trans. on Commun.*, Vol. 51, Issue 6, pp. 86-93, Dec. 2018.

[2] L. Liu and W. Yu, "Massive Connectivity With Massive MIMOPart I: Device Activity Detection and Channel Estimation," in *IEEE Trans. on Sig. Process.*, vol. 66, no. 11, pp. 2933-2946, 1 June, 2018.

[3] H. Jiang, D. Qu, J. Ding, and T. Jiang, "Multiple Preambles for High Success Rate of Grant-Free Random Access With Massive MIMO" in *IEEE Trans. on Wireless Commun.*, vol. 18, no. 10, pp. 4779-4789, Oct. 2019.

[4] P. Popovski, J.J. Neilsen, C. Stefanovic, E. Carvalho, E. Strom, K. F. Trillingsgaard, A. Bana, D. M. KIm, R. Kotaba, J. Park, and R. B. Sorensen "Wireless Access for Ultra-Reliable Low-Latency Communication: Principles and Building Blocks," in *IEEE Network*, vol. 32, no. 2, pp. 16-23, March-April 2018.

[5] R. Gribonval, B. Mailhe, H. Rauhut, K. Schnass and P. Vandergheynst, "Average Case Analysis of Multichannel Thresholding," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, Honolulu, HI, 2007, pp. II-853-II-856.