





# Hardware Implementation of Task-Based Quantization in Multiuser Signal Recovery

Xing Zhang , Member, IEEE, Haiyang Zhang , Member, IEEE, Nimrod Glazer , Member, IEEE, Oded Cohen, Eliya Reznitskiy, Shlomi Savariego, Moshe Namer, and Yonina C. Eldar , Fellow, IEEE

**Abstract**—Quantization plays a critical role in digital signal processing systems, allowing the representation of continuous-amplitude signals with a finite number of bits. However, accurately representing signals requires a large number of quantization bits, which causes severe cost, power consumption, and memory burden. A promising way to address this issue is task-based quantization. By exploiting the task information for the overall system design, task-based quantization can achieve satisfying performance with low quantization costs. In this work, we apply task-based quantization to multiuser signal recovery and present a hardware prototype implementation. The prototype consists of a tailored configurable combining board, and a software-based processing and demonstration system. Through experiments, we verify that with proper design, the task-based quantization achieves a reduction of 25 fold in memory by reducing from 16 receivers with 16 bits each to 2 receivers with 5 bits each, without compromising signal recovery performance.

**Index Terms**—Analog combiner, hardware implementation, multiuser signal recovery, task-based quantization.

## I. INTRODUCTION

PROCESSING and storing information that originates as analog signals involves converting this information to bits by analog-to-digital converters (ADCs) [1]. In conventional

receivers, the ADC is employed as a separate unit regardless of other parts of the system. ADCs typically sample at the Nyquist rate of the received signal and use high-resolution quantizers, so that sampling and quantization errors can be minimized [2]. However, since the power consumption of ADCs and the required storage memory grow with the sampling rate and quantization resolution, conventional ADCs pose great challenges to practical applications with high data rates. For example, in future 6G wireless communication systems, where hundreds or even thousands of antennas and millimeter-wave or subterahertz signaling are employed, it is expected that up to 1 Tb/s data rate will be achieved [3], [4]; in Internet of Things, machine-to-machine communication supports an enormous number of random access user equipments, which inevitably involves frequent data transmission and processing [5]. In such cases, the hardware implementation of high-resolution ADCs becomes a bottleneck. Therefore, more efficient sampling and quantization schemes are necessary.

Two prominent research directions to alleviate the power burden of ADCs are sub-Nyquist sampling [1], [6], [7] and low-resolution quantization [8], [9], [10], [11], [12]. Sub-Nyquist sampling aims to reduce the sampling rate by exploiting the underlying structure information of the signal [6]. For instance, the nature of finite rate of innovation signals has been exploited to reduce the sampling rate of received signals in ultrasound, radar, and cognitive radio [1], [7]. However, the sub-Nyquist sampling framework does not take into account the effect of quantization. As the power consumption of ADC increases in an exponential manner with the number of quantization bits, low-precision quantization has attracted great interest in recent years. Low-resolution quantization uses a few or even 1-bit to discretize the signal amplitude. It has been applied in various applications, such as massive multi-input multi-output (MIMO) communications [8], [9], radar [10], direction of arrival estimation [11], and spectrum sensing [12]. Compared with conventional high-resolution quantizers, significant rate reduction can be expected by using low-bit quantizers. However, compensation for the distortion induced in quantization is required in subsequent digital processing, which results in complicated information extraction in the digital domain and an overall system performance degradation.

To address these issues, the authors of [13] proposed task-based quantization. In task-based quantization systems, the quantizer is no longer a separate part, but codesigned with the analog and digital modules of the system under the objective

Manuscript received 31 January 2023; revised 28 May 2023 and 16 July 2023; accepted 14 August 2023. Date of publication 7 September 2023; date of current version 29 February 2024. This work was supported in part by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program under Grant 101000967, in part by the Israel Science Foundation under Grant 536/22, in part by the Many Igel Centre for Biomedical Engineering and Signal Processing, and in part by the Natural Science Research Start-up Foundation of Recruiting Talents of Nanjing University of Posts and Telecommunications under Grant NY222117. (Corresponding authors: Xing Zhang; Haiyang Zhang.)

Xing Zhang and Haiyang Zhang are with the School of Communication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210049, China (e-mail: xing\_zhang@njupt.edu.cn; haiyang.zhang@njupt.edu.cn).

Nimrod Glazer, Oded Cohen, Eliya Reznitskiy, Shlomi Savariego, and Yonina C. Eldar are with the Faculty of Math and Computer Science, Weizmann Institute of Science, Rehovot 7610001, Israel (e-mail: nimrod.glazer@weizmann.ac.il; oded.cohen@weizmann.ac.il; eliya.reznitskiy@weizmann.ac.il; shlomi.savariego@weizmann.ac.il; yonina.eldar@weizmann.ac.il).

Moshe Namer is with the Department of Electrical Engineering, Technion, Haifa 3200003, Israel (e-mail: namer@ee.technion.ac.il).

Color versions of one or more figures in this article are available at <https://doi.org/10.1109/TIE.2023.3310029>.

Digital Object Identifier 10.1109/TIE.2023.3310029

to minimize the task recovery error. By doing so, task-based quantization dramatically reduces the number of bits and thus the power consumption and storage burden of the system, without compromising the task recovery performance. The idea of task-based quantization has been applied to graph signal processing [14], channel estimation in massive MIMO communications [15], target identification in radar [16], and receiver design of dual function radar-communications [17]. Furthermore, Shlezinger et al. [18] proposed using deep networks to learn the task-based sampling and quantization process, and showed that the task-based method achieves performance comparable to operating with high sampling rates and fine resolution quantization, while operating with reduced overall bit rate. While Bernardo et al. [19] considered the design and analysis of task-based quantization system that are equipped with nonuniform scalar quantizers or that have inputs with unbounded support. Theoretical maturity of the concept suggests the need to demonstrate and evaluate the implementation of such systems in hardware, however, both the hardware prototype and the corresponding experiments remain to be investigated, which is the focus of this work. The main contributions of this work are as follows.

- 1) We apply task-based quantization to multiuser signal recovery in massive MIMO systems. In this setting, especially when the base station (BS) is equipped with large-scale antenna arrays, assigning each antenna with a high-resolution quantizer is impractical due to the hardware and power concerns. Considering the fact that the system task is to recover multiuser transmitted signals, rather than the received signals on all antennas, the idea of exploiting the task information to reduce quantization bits meets the practical concerns appropriately.
- 2) A hardware prototype, which consists of a tailored configurable combining board and a software-based processing and demonstration system, is presented. The analog combiner is a self-designed hardware that realizes a controllable analog combiner network, where both the gain and phase of each analog combining weight can be adjusted. The demonstration system is a MATLAB-based graphical user interface (GUI) that allows parameter controlling, data processing, and results displaying.
- 3) We verify through experiments that the task-based quantization outperforms the conventional task-ignorant one, i.e., the task-based quantization achieves a reduction of 25 fold in memory by reducing from 16 receivers with 16 bits each to 2 receivers with 5 bits each, without compromising signal recovery performance. These results mitigate the gap between the task-based quantization theory and its practical applications.

The rest of this article is organized as follows. Section II formulates the problem of task-based quantization for multiuser signal recovery and provides the theoretical results. Next, in Section III, the system architecture and each component of the hardware prototype are introduced in detail. Experimental results are provided in Section IV. Finally, Section V concludes this article.

*Notation:* Scalar quantities, column vectors, and matrices are denoted by lowercase letters,  $a$ , bold lowercase letters,  $\mathbf{a}$ ,

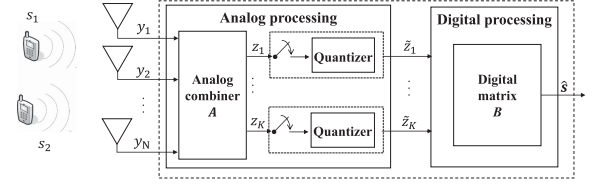


Fig. 1. Illustration of task-based quantization for multiuser signal recovery.

and bold uppercase letters,  $\mathbf{A}$ , respectively. The superscripts  $(\cdot)^T$ ,  $(\cdot)^{-1}$ , and  $(\cdot)^H$  are, respectively, the transpose, inverse, and Hermitian transpose operators. The symbol  $E[\cdot]$  represents statistical expectation,  $\|\cdot\|$  is the Euclidean norm,  $\mathcal{C}$  is the set of complex numbers, and  $\mathbf{I}_K$  is the  $K \times K$  identity matrix. We use  $a^+$  to denote  $\max(a, 0)$ , and  $\lfloor \cdot \rfloor$  to denote rounding down to the next smaller integer.

## II. TASK-BASED QUANTIZATION FOR MULTIUSER SIGNAL RECOVERY

In this section, we provide a mathematical description of multiuser signal recovery under task-based quantization. In particular, we begin by introducing the system model of task-based quantization for multiuser signal recovery in Section II-A, followed by theoretical results in Section II-B.

### A. System Model

As shown in Fig. 1, we consider a multiuser MIMO scenario where the BS equipped with  $N$  antennas serves  $K$  single-antenna user terminals (UTs). The received signal at the BS, denoted by  $\mathbf{y}$ , can be expressed as

$$\mathbf{y} = \mathbf{H}\mathbf{s} + \mathbf{v} \quad (1)$$

where  $\mathbf{s} \in \mathcal{C}^{K \times 1}$  denotes the transmitted signals of  $K$  UTs;  $\mathbf{v} \in \mathcal{C}^{N \times 1}$  represents additive white Gaussian noise;  $\mathbf{H} \in \mathcal{C}^{N \times K}$  denotes the wireless channel, with its  $k$ th column representing the channel between user  $k$  and the antenna array, given by [20], [21]

$$\mathbf{h}_k = g_k e^{-j2\pi \frac{f_c r_k}{c}} \mathbf{a}(\theta_k). \quad (2)$$

Here,  $g_k$  denotes the path gain, where without loss of generality, we assume only the line of sight path exists,  $c$  is the speed of light,  $f_c$  is the carrier frequency,  $r_k$  and  $\theta_k$  are, respectively, the distance and angle of arrival of the  $k$ th user. The vector  $\mathbf{a}(\theta_k)$  is the steering vector, given by

$$\mathbf{a}(\theta_k) = \left[ 1, e^{j\pi \sin \theta_k}, \dots, e^{j\pi(N-1)\sin \theta_k} \right]^T. \quad (3)$$

We assume the channel is quasi-static over the signal transmission time, so that  $\mathbf{H}$  can be estimated by pilots and is assumed to be known for the task of recovering  $\mathbf{s}$ .

In conventional quantization systems, ADCs are only used to discretize the received signal  $\mathbf{y}$ . The task of recovering  $\mathbf{s}$  is performed separately in the digital domain. By contrast, task-based quantization proposed in [13] jointly designs the overall analog and digital system to estimate  $\mathbf{s}$ . Under this principle, the

multiuser signal recovery system, as shown in Fig. 1, consists of the following three parts: an analog combiner matrix  $\mathbf{A}$ ,  $K$  scalar quantizers, and a digital processing module. The analog combiner is introduced to reduce the dimensionality of the quantizer input based on the fact that the number of users is much smaller than the number of antennas, so that the needed number of quantizers can be reduced. Then, scalar quantizers with limited resolution are used to digitize the combined signal and finally, a digital matrix is employed to recover the user signal  $\mathbf{s}$ . Mathematically, the received signal  $\mathbf{y}$  is first projected to a  $K \times 1$  vector  $\mathbf{z}$  by using the analog combiner  $\mathbf{A}$ , i.e.,

$$\mathbf{z} = \mathbf{A}\mathbf{y} = \mathbf{A}\mathbf{H}\mathbf{s} + \mathbf{A}\mathbf{v}. \quad (4)$$

Then, each entry of  $\mathbf{z}$  is sampled and quantized using scalar quantizers with dynamic range  $\gamma$  and resolution  $\tilde{M}_K \triangleq \lfloor M^{1/K} \rfloor$ . The symbol  $M$  is the overall number of quantization levels, which represents the memory requirement of the system and is also directly related to the ADC power consumption. When the input is inside the dynamic range of the quantizer, the output can be written as the sum of the input and an additive zero-mean white noise signal according to the theory of dithered quantization [22], that is

$$\tilde{\mathbf{z}} = \mathbf{A}\mathbf{H}\mathbf{s} + \mathbf{A}\mathbf{v} + \mathbf{e} \quad (5)$$

where  $\mathbf{e}$  is the quantization noise with covariance  $\frac{\Delta^2}{2}\mathbf{I}_K$ . The symbol  $\Delta$  denotes the quantization spacing defined as  $\Delta = \frac{2\gamma}{M_K}$ . When  $\tilde{M}_K$  is given, the value of  $\gamma$  determines the quantization spacing, and therefore, the variance of the quantization noise.

In the digital domain, the estimation of  $\mathbf{s}$ , denoted as  $\hat{\mathbf{s}}$ , is obtained as the output of the digital processing module  $\mathbf{B}$ , yielding

$$\hat{\mathbf{s}} = \mathbf{B}\tilde{\mathbf{z}}. \quad (6)$$

The problem now is to jointly design the analog combiner  $\mathbf{A}$ , the dynamic range  $\gamma$ , and the digital processing matrix  $\mathbf{B}$ , so that the mean square error (MSE) of the task estimate can be minimized. Mathematically, we have the following optimization problem:

$$\min_{\mathbf{A}, \gamma, \mathbf{B}} E[|\mathbf{s} - \hat{\mathbf{s}}|^2]. \quad (7)$$

## B. Theoretical Results

According to the orthogonality principle, the MSE in (7),  $E[|\mathbf{s} - \hat{\mathbf{s}}|^2]$ , can be re-expressed as

$$E[|\mathbf{s} - \hat{\mathbf{s}}|^2] = E[|\mathbf{s} - \tilde{\mathbf{s}}|^2] + E[|\tilde{\mathbf{s}} - \hat{\mathbf{s}}|^2] \quad (8)$$

where  $\tilde{\mathbf{s}}$  is the linear minimum mean square error (LMMSE) estimate of  $\mathbf{s}$  from  $\mathbf{y}$ , that is,  $\tilde{\mathbf{s}} = \mathbf{\Gamma}\mathbf{y}$ , with  $\mathbf{\Gamma}$  denoting the LMMSE estimation matrix. Note that the first term in the aforementioned equation is independent of  $\hat{\mathbf{s}}$ . The optimization problem in (7) can, thus, be equivalently replaced by

$$\min_{\mathbf{A}, \gamma, \mathbf{B}} E[|\tilde{\mathbf{s}} - \hat{\mathbf{s}}|^2] \quad (9)$$

which is the same as [13]. Therefore, in the following, we directly provide the obtained optimization results and omit the proof.

Let  $\Sigma_{\mathbf{y}}$  be the covariance matrix of the received signal  $\mathbf{y}$ , and  $w_l$ ,  $l = 1, \dots, K$  the dither signal added to the input of the  $l$ th quantizer. Let  $\mathbf{A}^\circ$  and  $\mathbf{B}^\circ$  the optimal analog and digital processing matrices that achieve the minimal MSE distortion. Then, we have the following results [13].

*Theorem 1:* For any analog combining matrix  $\mathbf{A}$  and dynamic range  $\gamma$  such that  $\Pr(|(\mathbf{A}\mathbf{y})_l + w_l| > \gamma) = 0$ , namely, the quantizers operate within their dynamic range with probability one, the digital processing matrix which minimizes the MSE is given by

$$\mathbf{B}^\circ(\mathbf{A}) = \mathbf{\Gamma}\Sigma_{\mathbf{y}}\mathbf{A}^H \left( \mathbf{A}\Sigma_{\mathbf{y}}\mathbf{A}^H + \frac{2\gamma^2}{\tilde{M}_K^2 \cdot K} \mathbf{I}_K \right)^{-1}. \quad (10)$$

*Theorem 2:* For the hardware-limited quantization system based on the model depicted in Fig. 1, the optimal analog combining matrix is given by  $\mathbf{A}^\circ = \mathbf{U}_\mathbf{A} \mathbf{\Lambda}_\mathbf{A} \mathbf{V}_\mathbf{A}^H \Sigma_{\mathbf{y}}^{-1/2}$ , where

- 1)  $\mathbf{V}_\mathbf{A} \in \mathcal{C}^{N \times N}$  is the right singular vectors matrix of  $\tilde{\mathbf{\Gamma}} \triangleq \mathbf{\Gamma}\Sigma_{\mathbf{y}}^{1/2}$ ;
- 2)  $\mathbf{\Lambda}_\mathbf{A} \in \mathcal{C}^{K \times N}$  is a diagonal matrix with diagonal entries

$$(\mathbf{\Lambda}_\mathbf{A})_{i,i}^2 = \frac{2\kappa_p}{\tilde{M}_K^2 \cdot K} \left( \zeta \cdot \lambda_{\tilde{\mathbf{\Gamma}},i} - 1 \right)^+$$

where  $\kappa_p = \eta^2 \left(1 - \frac{2\eta^2}{3\tilde{M}_K^2}\right)^{-1}$  with  $\eta$  denoting a constant that is set to guarantee that the quantizer operates within the dynamic range [13],  $\{\lambda_{\tilde{\mathbf{\Gamma}},i}\}$  are singular values of  $\tilde{\mathbf{\Gamma}}$  arranged in a descending order, and  $\zeta$  is chosen such that

$$\frac{2\kappa_p}{\tilde{M}_K^2 \cdot K} \sum_{i=1}^K \left( \zeta \cdot \lambda_{\tilde{\mathbf{\Gamma}},i} - 1 \right)^+ = 1.$$

- 3)  $\mathbf{U}_\mathbf{A} \in \mathcal{C}^{K \times K}$  is a unitary matrix which guarantees that  $\mathbf{U}_\mathbf{A} \mathbf{\Lambda}_\mathbf{A} \mathbf{\Lambda}_\mathbf{A}^H \mathbf{U}_\mathbf{A}^H$  has identical diagonal entries.

The dynamic range of the quantizer is given by

$$\gamma^2 = \frac{\eta^2}{K} \left( 1 - \frac{2\eta^2}{3\tilde{M}_K^2} \right)^{-1} \quad (11)$$

and the resulting minimal achievable distortion is

$$E[|\tilde{\mathbf{s}} - \hat{\mathbf{s}}|^2] = \sum_{i=1}^K \frac{\lambda_{\tilde{\mathbf{\Gamma}},i}^2}{\left( \zeta \cdot \lambda_{\tilde{\mathbf{\Gamma}},i} - 1 \right)^+ + 1}. \quad (12)$$

Next, we analyze the computational complexity of calculating  $\mathbf{B}^\circ(\mathbf{A})$  in Theorem 1 and  $\mathbf{A}^\circ$  in Theorem 2. The dominant complexity of computing  $\mathbf{B}^\circ(\mathbf{A})$  comes from the matrix inversion operation, which is  $\mathcal{O}(K^3)$ . While the computation of  $\mathbf{A}^\circ$  consists of four parts: the calculation of  $\mathbf{V}_\mathbf{A} \in \mathcal{C}^{N \times N}$  requires the singular value decomposition operation, whose computational complexity is  $\mathcal{O}(N^3)$ . The optimal  $\zeta$  in  $\mathbf{\Lambda}_\mathbf{A}$  can be found using the simple bisection search method, whose complexity is  $\mathcal{O}(\log(\frac{\Delta\zeta}{\epsilon}))$ , where  $\Delta\zeta$  and  $\epsilon$  represent the initial interval size and desired accuracy, respectively. The computational complexity of matrix inversion for matrix  $\Sigma_{\mathbf{y}}$  is also  $\mathcal{O}(N^3)$ .  $\mathbf{U}_\mathbf{A}$  can be obtained from  $\mathbf{\Lambda}_\mathbf{A}$  using the unitary transformation operation with the complexity of  $\mathcal{O}(K^3 + K^2)$ . Therefore, the total



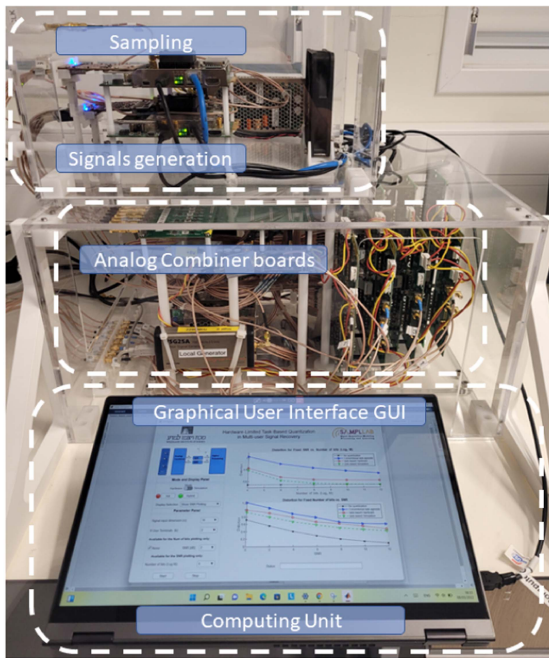


Fig. 2. Task-based quantization system.

computational complexity of  $\mathbf{A}^\circ$  in Theorem 2 is approximately  $\mathcal{O}(2N^3 + K^3 + K^2 + \log(\frac{\Delta\zeta}{\epsilon}))$ .

In our prototype, we configure the analog combiner according to Theorem 2, and the dynamic range of the scalar quantizer based on (11). The calculated matrix  $\mathbf{B}^\circ$  in (10) is used for the task vector recovery in the digital domain. Details of the hardware implementation are discussed in the following section.

### III. HARDWARE IMPLEMENTATION

This section elaborates on the system architecture of the hardware prototype, which realizes task-based quantization for multiuser signal recovery detailed in the previous section. We first present the high-level system architecture in Section III-A. The concrete structure of each component is provided in Section III-B, and the design challenges are detailed in Section III-C.

#### A. High-Level Architecture

Fig. 2 shows our hardware board, which consists of five main blocks: GUI, signal generator, Analog combiner board, sampling, and computing center. Details of the employed hardware components are presented in Table I, and an overview of the information flow of the system in the hardware implementation is provided in Fig. 3. As shown in this figure, the system model described in Section II-A is implemented by three steps: first, the MATLAB numerically simulates the process of signal transmitting, passing through the channel and being received by antennas at the BS, see (1); then, the hardware board converts the digital data generated by MATLAB to analog waveform signals, and implements analog combining, i.e., (4), and sampling; finally, the low-resolution quantization and digital-domain task

TABLE I  
LIST OF HARDWARE COMPONENTS

Component	Model number	Make
FPGA	Xilinx VC707	Texas Instruments
DAC	FMC204 FPGA Card	Abaco Systems
ADC	FMC168 FPGA Card	Abaco Systems
Local oscillator	VSG25A	Signal Hound

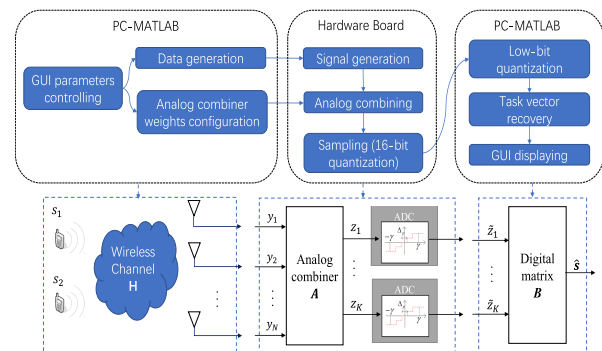


Fig. 3. Overview of the information flow of the system in the experimental setup.

recovery, i.e., (6), are again performed by MATLAB. The major building components are as follows.

**1) GUI:** The GUI is used for controlling the system parameters, which allows the user to configure the experimental setup in a user-friendly environment. The main controllable parameters include the number of UTs, receiving antennas, quantization bits, and the SNRs of the received signals. Based on these parameters, the MATLAB running on the computing unit generates input data for 16-channel digital-to-analog converters (DACs) that are located at the analog combiner board (details in Section III-A3), and the optimal weights for the analog combiner configuration.

**2) Signal Generator:** The digital data generated by MATLAB is then fed to a field programmable gate array (FPGA) board with 16 transmit channels DACs to generate analog waveform signals. This process is adopted to mimic the real-world receiving signals at the BS.

**3) Analog Combiner Board:** The 16 baseband (BB) analog signals from the DAC are next fed into the analog combiner board, as illustrated in Fig. 4. To transmit them in the desired frequency, they are up-converted to 2.3 GHz by 16 dedicated modulators. Then, the signal of each channel is passed through a 4-way power divider (splitter), yielding 64 analog RF signals in total. The 64 signals are then fed into a 4-combiner boards. Each combiner board is fed by 16 channels and has a single output. The combiner board is controlled by an analog vector multipliers device designed to control a signal's gain and phase. The overall process is illustrated in Fig. 5. In this way, the tailored analog

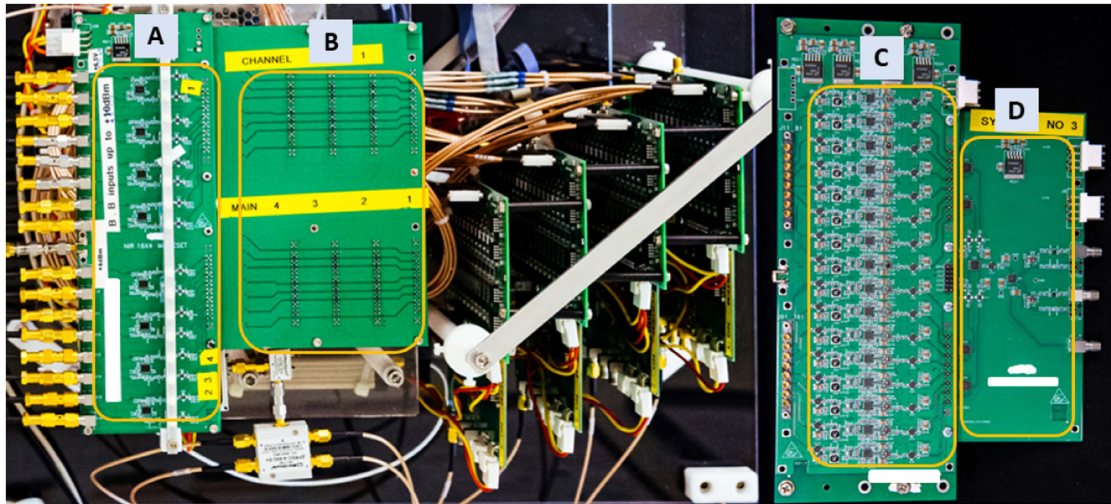


Fig. 4. Analog combiner hardware board.

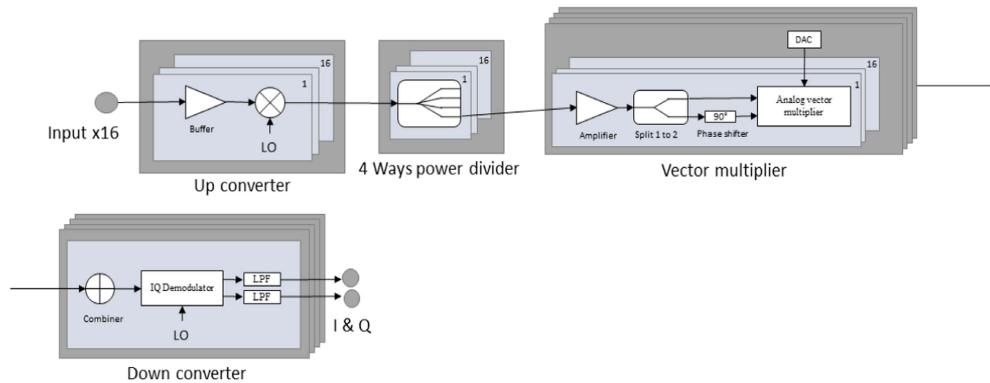


Fig. 5. Analog combiner schematics.

combiner board converts 16-channel signals to 4, implementing the function of the aforementioned analog combiner matrix A.

**4) Sampling:** The outputs (both  $I$  and  $Q$ ) of the analog combiner board are fed into a sampling board. The four analog signals are down-converted from 2.3 GHz to 20 MHz, and are converted to digital signals by using 4DSP FMC168 16-bit digitizer card.

**5) Computing Center:** The four digital streams are then transferred to the MATLAB application on the computing center. The MATLAB mimics a digital low-bit quantization and then recovers multiuser signals in the digital domain. Finally, the results are displayed on the GUI to demonstrate the signal recovery performance of the task-based hardware prototype.

### B. Details of Each Block

**1) Waveform Generation:** The 16 digital BB signals generated by the host application are transferred to the FPGA board in real time by an Ethernet cable. The FPGA board generates the corresponding analog BB signals waveform with a maximal frequency range of 100 MHz.

**2) Analog Combiner:** The analog combiner board is a self-designed dedicated hardware that realizes a controllable analog combiner network. As shown in Fig. 4, the board consists of the following four parts.

- Up-conversion:** The 16 input complex BB signals, whose maximal frequency range is 100 MHz, are up-converted to RF signals using a VSG25 A vector signal generator. By up-conversion, the RF signals can represent the passband signals observed at the BS.
- Passband signals splitting:** The analog passband signal of each channel is split into four. In the considered setting here, we have 64 analog RF signals in total, which are combined for further processing. The board can support 4 RF-chain processing. Since the number of users is set as 2 in this experiment, we only use 2 of them to process the output RF signals.
- Parameter generation and configuration for the combiner:** Each split signal is fed into an amplifier, split again into two signals with a 90-degree offset. The two signals then enter into an ADL5390 analog vector multiplier. The

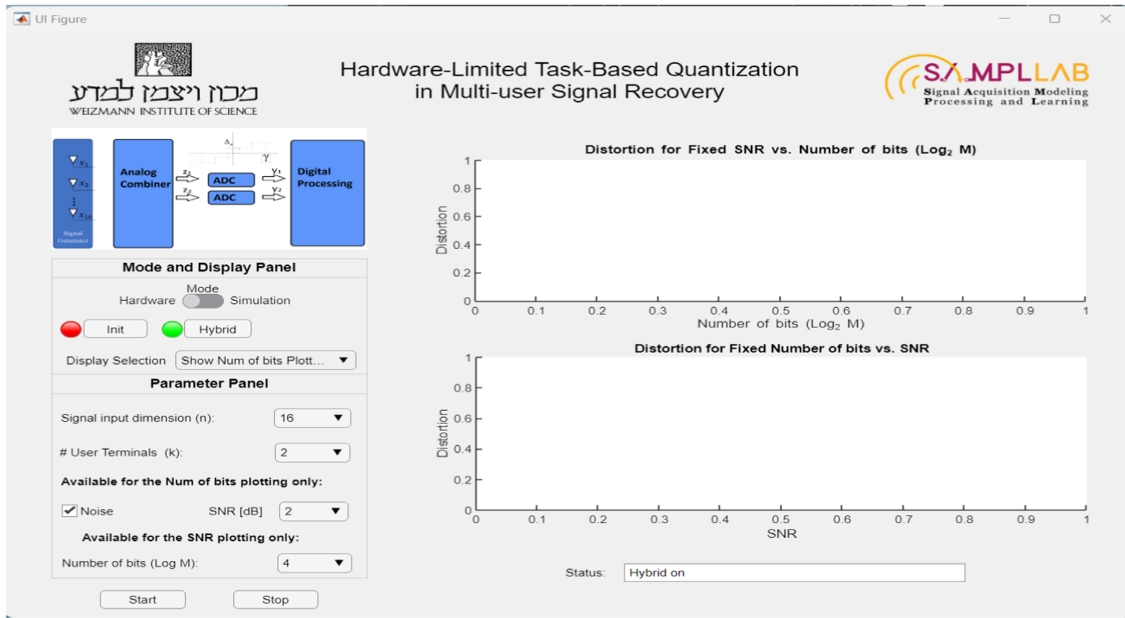


Fig. 6. Overview of the GUI.

analog vector multiplier implements the phase and gain of each analog combining weight, which is applied to combine the input signal. The applied weights are determined by the output dc level of an AD5674 octal 12-bit DACs with serial load capabilities, which receives control commands via Arduino Nano microcontroller device to configure the analog combining weights. The usage of controllable gains and phases requires a calibration stage when the interconnections are established, to guarantee that the configured weights are correctly translated into the desired phase and gain values.

- d) Summing up of the incoming signals and down-conversion: The final step is summing the 16 output signals of each group after weighting, to obtain a combined passband signal. The signal is then down-converted by the same local oscillator that is employed for up-conversion, and filtered to BB with a maximum 100 MHz bandwidth.

**3) Quantization:** The four output signals are forwarded to be sampled by the 4DSP FMC168 16-bit digitizer card. However, in task-based quantization, it is expected to use low-bit quantizers. We here use software simulation to mimic the hardware implementation of such a scalar quantizer defined as

$$q(x) \triangleq \begin{cases} \Delta \left( \lfloor \frac{x}{\Delta} \rfloor + \frac{1}{2} \right), & \text{for } |x| < \gamma \\ \text{sign}(x) \left( \gamma - \frac{\Delta}{2} \right), & \text{else} \end{cases} \quad (13)$$

where  $x$  is the input signal,  $\Delta = \frac{2\gamma}{\tilde{M}_K}$  represents the quantization spacing, with the variable  $\tilde{M}_K$  varied in the experiments for different number of bits. The symbol  $\text{sign}(\cdot)$  denotes the signum function.

**4) Software (Digital Processing):** The software part consists of two components: a computing center running the MATLAB-based host application, and a GUI-based control and display interface.

The computing center is a 64-bit computer with 8 CPU cores and 16 GB RAM running the MATLAB-based host application. The application is responsible for generating the digital BB signal, computing the optimal analog and digital processing matrices as detailed in Theorems 1 and 2, computing the dynamic range of the quantizer, and post-processing the digital output to recover the task vector.

The display part of the GUI presents the experiment results in two modes: the MSE distortion with respect to the number of bits, or SNR, as shown in Fig. 6. The control part provides a way for users to interact with the experiment setup, that is, it allows users to change the parameters used in the experiment. The main controllable parameters include the dimensionality of the received signal and the task vector, the SNR level for plotting MSE distortion versus the number of bits, and the number of bits for plotting MSE distortion versus SNR. Details of the supported parameter combinations are summarized in Table II.

### C. Design Challenges

As we detailed in Section III-B2, the phase and gain of each analog combining weight are determined by the amplitudes of in-phase and quadrature-phase output direct current created by AD5674 octal 12-bit DACs. In our case, we have  $16 \times 4 = 64$  RF signals that need to be weighted. Ideally, if all RF chains operate within the linear dynamic range of the device, the combination of the 64 RF signals of the 4 boards will result in an accurate summation as expected. However, it is not the case in practical experiments.

To address this issue, we introduced a calibration process that scanned through the amplitude and phase of each RF-chain and performed relevant modifications. This process is done by setting the DAC value for adjusting the in-phase and quadrature-phase amplitudes of each signal. As shown in Fig. 7 which



TABLE II  
CONTROLLABLE PARAMETERS SUPPORTED BY GUI

Working mode	Simulation		Hardware	
Curve display mode	NMSE versus number of bits	NMSE versus SNR	NMSE versus number of bits	NMSE versus SNR
Number of UTs	$K = 2, 4, 8$		$K = 2$	
Number of receiving antennas in the BS	$N = 4, 8, 16, 60, 120$		$N = 16$	
Noise	SNR = 2, 4, 6, 8, 10		SNR = 2, 4, 6, 8, 10	
Number of bits ( $\log_2 M$ )		4, 8, 12, 16, 20		4, 8, 12, 16, 20

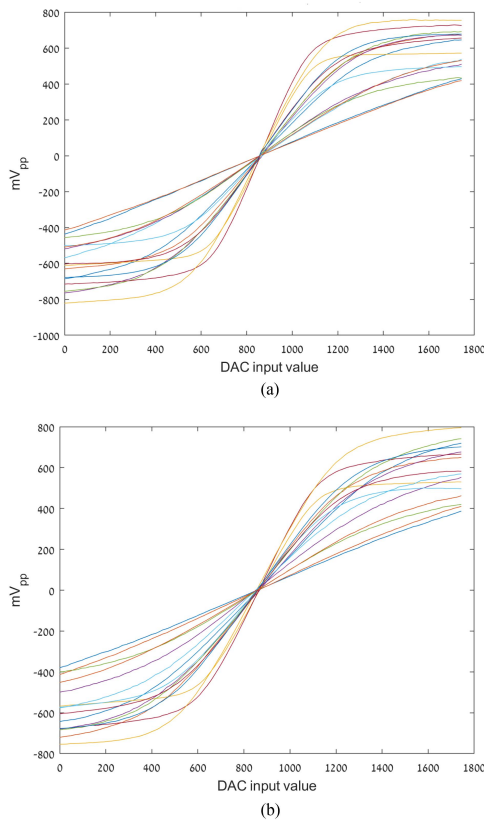


Fig. 7. Illustration of the calibrated signals of the 16 channels of Board 1, (a) In-phase, (b) Quadrature-phase.

illustrates the calibration results of the 16 channels of Board 1, with Fig. 7(a) presenting the in-phase signals and Fig. 7(b) the quadrature-phases signals, we see that after calibration, at the DAC steps 700–1100, the output voltage varies nearly linearly with the input DAC step. Therefore, we chose this range in order to program the DAC to achieve the desired combining outputs. In our experiments, the calibration process is performed iteratively until the outputs of the 16 RF chains falls into a small concentrated area. In particular, as shown in Fig. 8, at each phase point, the Euclidean distance of the 16 signals from its center is comparatively small after iterations.

#### IV. HARDWARE RESULTS

In this section, hardware experiments are carried out to evaluate the performance of task-based quantization in multiuser signal recovery. We consider the case where the number of users is  $K = 2$ , and the number of antennas at the BS is  $N = 16$ . The transmitted signal from the two users obeys zero-mean and

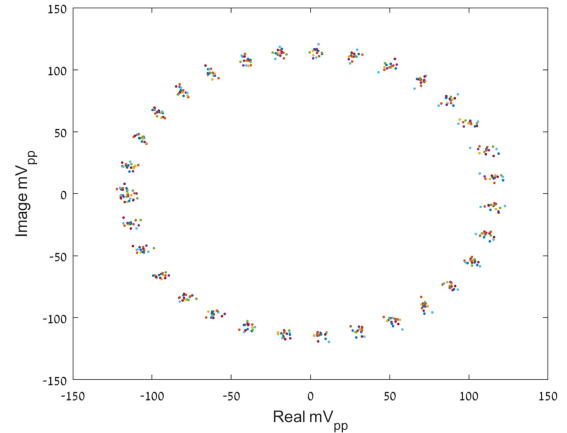


Fig. 8. Illustration of the phase distribution of the calibrated signals.

unit variance Gaussian distribution, and the channel is generated based on (2) with  $L = 3$  paths for each user. All the results are obtained by averaging 2000 experiments.

As a comparison, task-agnostic vector quantization results are included. Different from scalar quantizers which operate on a scalar input, vector quantizers have a multivariate input. Therefore, vector quantization cannot be implemented using practical serial scalar ADCs. Here, we employ simulated task-agnostic vector quantization as a comparison since it represents the best system one can construct when the quantizer is designed separately from the task [13]. Furthermore, the ideal case where no quantization is imposed on the sampled signal is also provided as a benchmark. The GUI provides two modes: simulation mode and hardware mode. In the simulation mode, the combining, sampling, and quantization of the received signal are all performed by software, i.e., MATLAB. In hardware mode, the combining and sampling are performed by the hardware board. We set SNR = 2 dB in the case of displaying distortion versus the number of bits, and the number of total used equals 4 for plotting distortion versus SNRs. The results are shown in Figs. 9 and 10 where both the simulation and hardware results are provided in the same figure.

From these results, we see that task-based quantization significantly outperforms task-agnostic vector quantization, and can approach the optimal performance with the increase of  $M$ . In particular, when each quantizer is assigned more than five bits, i.e.,  $\log_2 M \geq 5K$ , the quantization error becomes negligible. This suggests that by exploiting prior knowledge of the task, and by properly designing the overall system, task-based quantization can achieve satisfying performance with a much smaller

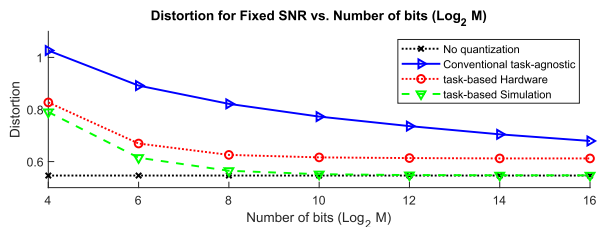


Fig. 9. MSE distortion versus the number of total bits.

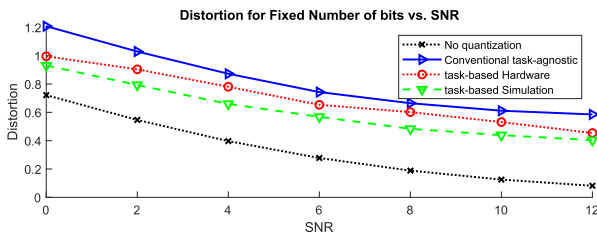


Fig. 10. MSE distortion versus the SNRs.

number of bits, i.e., from 16 receivers with 16 bits each to 2 receivers with 5 bits each. Furthermore, the hardware results agree with the simulated ones, with only a small performance gap caused by imperfect hardware calibration and hardware noise, verifying the effectiveness of the task-based quantization hardware prototype.

## V. CONCLUSION

With the increase of data rate, conventional ADCs, which sample at the Nyquist rate and use high-resolution quantizers face challenges in storage and power consumption. To reduce quantization bits, task-based quantization has been proposed by exploiting the underlying task for the system design. In this work, we presented the application of task-based quantization in multiuser signal recovery and provided a hardware implementation. The prototype consists of a tailored configurable analog combiner board and a software-based processing and demonstration system. Experimental results illustrate the superiority of task-based quantization over conventional ADCs, mitigating the gap between the theory and its practical application. We envision that the hardware prototype and experiments presented in this work will promote the development of the task-based quantization in practical communication systems. In this sense, it is promising to chipify the hardware board and employ cost-effective elements considering the space and cost constraints of communication transceivers in practice, which remains to be investigated in the future.

## REFERENCES

- [1] Y. C. Eldar, *Sampling Theory: Beyond Bandlimited Systems*. Cambridge, U.K.: Cambridge Univ. Press, 2015.
- [2] M. Yang, H. Huang, Z. Liu, C. Cai, Y. Wang, and J. Yang, "Sine approximation-based comparison method for determining the phase-frequency characteristics of analog-to-digital converters," *IEEE Trans. Ind. Electron.*, vol. 71, no. 2, pp. 2142–2145, Feb. 2024.

- [3] N. Rajatheva et al., "White paper on broadband connectivity in 6G," 2020, *arXiv:2004.14247*.
- [4] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y. Zhang, "The roadmap to 6G: AI empowered wireless networks," *IEEE Commun. Mag.*, vol. 57, no. 8, pp. 84–90, Aug. 2019.
- [5] K. Lai, J. Lei, Y. Deng, L. Wen, G. Chen, and W. Liu, "Analyzing uplink grant-free sparse code multiple access system in massive IoT networks," *IEEE Internet Things J.*, vol. 9, no. 7, pp. 5561–5577, Apr. 2022.
- [6] A. Kipnis, A. Goldsmith, and Y. C. Eldar, "Analog-to-digital compression: A new paradigm for converting signals to bits," *IEEE Signal Process. Mag.*, vol. 35, no. 3, pp. 16–39, May 2018.
- [7] D. Cohen, S. Tsiper, and Y. C. Eldar, "Analog to digital cognitive radio: Sampling, detection and hardware," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 137–166, Jan. 2018.
- [8] A. B. Fernandes, Z. Shao, L. T. Landau, and R. C. Lamare, "Multiuser-MIMO systems using comparator network-aided receivers with 1-bit quantization," *IEEE Trans. Commun.*, vol. 71, no. 2, pp. 908–921, Feb. 2023.
- [9] X. Song, S. Ma, P. Neuhaus, W. Wang, X. Gao, and G. Fettweis, "On robust millimeter wave line-of-sight MIMO communications with few-bit ADCs," *IEEE Trans. Wireless Commun.*, vol. 21, no. 12, pp. 11164–11178, Dec. 2022.
- [10] A. Ameri, A. Bose, J. Li, and M. Soltanalian, "One-bit radar processing with time-varying sampling thresholds," *IEEE Trans. Signal Process.*, vol. 67, no. 20, pp. 5297–5308, Oct. 2019.
- [11] S. Sedighi, M. B. Shankar, M. Soltanalian, and B. Ottersten, "DoA estimation using low-resolution multi-bit sparse array measurements," *IEEE Signal Process. Lett.*, vol. 28, pp. 1400–1404, Jun. 2021.
- [12] A. Ali and W. Hamouda, "Power-efficient wideband spectrum sensing for cognitive radio systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 4, pp. 3269–3283, Apr. 2018.
- [13] N. Shlezinger, Y. C. Eldar, and M. R. D. Rodrigues, "Hardware-limited task-based quantization," *IEEE Trans. Signal Process.*, vol. 67, no. 20, pp. 5223–5238, Oct. 2019.
- [14] P. Li, N. Shlezinger, H. Zhang, B. Wang, and Y. C. Eldar, "Graph signal compression by joint quantization and sampling," *IEEE Trans. Signal Process.*, vol. 70, pp. 4512–4527, Sep. 2022.
- [15] N. Shlezinger, Y. C. Eldar, and M. R. D. Rodrigues, "Asymptotic task-based quantization with application to massive MIMO," *IEEE Trans. Signal Process.*, vol. 67, no. 15, pp. 3995–4012, Aug. 2019.
- [16] F. Xi, N. Shlezinger, and Y. C. Eldar, "BiLiMO: Bit-limited MIMO radar via task-based quantization," *IEEE Trans. Signal Process.*, vol. 69, pp. 6267–6282, Sep. 2021.
- [17] D. Ma, N. Shlezinger, T. Huang, Y. Liu, and Y. C. Eldar, "Bit constrained communication receivers in joint radar communications systems," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 8243–8247.
- [18] N. Shlezinger, A. Amar, B. Luitjen, R. J. Sloun, and Y. C. Eldar, "Deep task-based analog-to-digital conversion," *IEEE Trans. Signal Process.*, vol. 70, pp. 6021–6034, Dec. 2022.
- [19] N. Bernardo, J. Zhu, Y. C. Eldar, and J. Evans, "Design and analysis of hardware-limited non-uniform task-based quantizers," *IEEE Trans. Signal Process.*, vol. 71, pp. 1551–1562, Apr. 2023.
- [20] X. Yand, F. Cao, M. Matthaiou, and S. Jin, "On the uplink transmission of extra-large scale massive MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 69, no. 12, pp. 15229–15243, Dec. 2020.
- [21] Z. He, X. Yuan, and L. Chen, "Super-resolution channel estimation for massive MIMO via clustered sparse Bayesian learning," *IEEE Trans. Veh. Technol.*, vol. 68, no. 6, pp. 6156–6160, Jun. 2019.
- [22] R. M. Gray and T. G. Stockholm, "Dithered quantizers," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 805–812, May 1993.



**Xing Zhang** (Member, IEEE) received the B.S. and Ph.D. degrees in information engineering and in information and communication engineering from the School of Information Science and Engineering, Southeast University, Jiangsu, China, in 2015 and 2021, respectively.

From 2021 to 2022, she was a Postdoctoral Research Fellow with the Weizmann Institute of Science, Rehovot, Israel. She is currently an Associate Professor with the School of Communication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, China. Her research interests include signal processing and 6G communications.





**Haiyang Zhang** (Member, IEEE) received the B.S. degree in communication engineering from Lanzhou Jiaotong University, Lanzhou, China, in 2009, the M.S. degree in information and communication engineering from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2012, and the Ph.D. degree in information and communication engineering from Southeast University, Nanjing, China, in 2017.

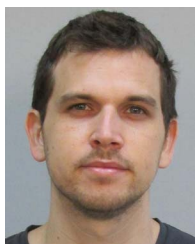
He is currently a Professor with the School of Communications and Information Engineering, Nanjing University of Posts and Telecommunications. From 2017 to 2020, he was a Postdoctoral Research Fellow with the Singapore University of Technology and Design, Singapore. From 2020 to 2022, he was a Postdoctoral Research Fellow with the Weizmann Institute of Science, Rehovot, Israel. His research interests include 6G near-field MIMO communications, deep learning, and sampling theory.

Dr. Zhang was awarded the FGS Prize for outstanding research achievements at Weizmann Institute of Science.



**Nimrod Glazer** (Member, IEEE) received the B.Sc. degree in electrical, electronics, and communications engineering from the Ben Gurion University of the Negev, Beersheba, Israel, in 1996.

From 1996 to 2020, he held various leadership positions with Motorola in Engineering, Innovation, and Research and Development. Since 2020, he has been the SAMPL Lab Manager, headed by Prof. Yonina Eldar, with the Mathematics and Computer Science Department, Weizmann Institute of Science, Rehovot, Israel.



**Oded Cohen** received the B.Sc. degree in electrical and computer engineering from the Ben Gurion University of the Negev, Be'er Sheva, Israel, in 2019.

He is currently a Software Engineer with SAMPL Lab, headed by Prof. Yonina Eldar, Mathematics and Computer Science Department, Weizmann Institute of Science, Rehovot, Israel.



**Eliya Reznitskiy** received the B.Sc. degree in electrical and electronics engineering from the Bar Ilan University, Ramat Gan, Israel, in 2022.

From 2020 to 2022, he worked on hardware implementations with SAMPL Lab, headed by Prof. Yonina Eldar, Mathematics and Computer Science Department, Weizmann Institute of Science, Rehovot, Israel. Since mid-2022, he has been a Chip Design Engineer with the Israel-based Samsung Research and Development Center.



**Shlomi Savariego** received the B.Sc. degree in electrical engineering from Tel Aviv University, Tel Aviv, Israel, in 1991.

From 1991 to 2020, he held various positions in the industry's real-time engineering research and development department. Since 2020, he has been the Real-Time Embedded and Ultrasound Engineer with SAMPL Lab, headed by Prof. Yonina Eldar, Mathematics and Computer Science Department, Weizmann Institute of Science, Rehovot, Israel.



**Moshe Namer** received the B.Sc. degree in electrical and communication engineering from the Technion—Israel Institute of Technology, Haifa, Israel, in 1984.

From 1984 to 2010, he was an Engineer with Communication Lab, Electrical Engineering Department, Technion, leading analog and RF circuits student projects in hardware implementations. Since 2010, he has been active with SAMPL Lab, headed by Prof. Yonina Eldar, working on hardware demo implementations.



**Yonina C. Eldar** (Fellow, IEEE) received the B.Sc. degree in physics and the B.Sc. degree in electrical engineering from Tel-Aviv University, Tel-Aviv, Israel, 1995 and 1996, respectively, and the Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, in 2002. Her research interests include statistical signal processing, sampling theory and compressed sensing, learning and optimization methods, and their applications to biology, medical imaging and optics.

She is currently a Professor with the Department of Mathematics and Computer Science, Weizmann Institute of Science, Rehovot, Israel. She was previously a Professor with the Department of Electrical Engineering, Technion, where she held the Edwards Chair in Engineering. She is also a Visiting Professor with MIT, a Visiting Scientist with the Broad Institute, and an Adjunct Professor with Duke University, and was a Visiting Professor with Stanford. She is a member of the Israel Academy of Sciences and Humanities (elected 2017) and a EURASIP Fellow. She is also the author of the book *Sampling Theory: Beyond Bandlimited Systems* (Cambridge University Press, 2015) and coauthor of four other books published by Cambridge University Press.

Dr. Eldar was the recipient of many awards for excellence in research and teaching, including the IEEE Signal Processing Society Technical Achievement Award (2013), the IEEE/AESS Fred Nathanson Memorial Radar Award (2014), and the IEEE Kiyo Tomiyasu Award (2016). She was a Horev Fellow of the Leaders in Science and Technology program with the Technion and an Alon Fellow. She was also the recipient of the Michael Bruno Memorial Award from the Rothschild Foundation, the Weizmann Prize for Exact Sciences, the Wolf Foundation Krill Prize for Excellence in Scientific Research, the Henry Taub Prize for Excellence in Research (twice), the Hershel Rich Innovation Award (three times), the Award for Women with Distinguished Contributions, the Andre and Bella Meyer Lectureship, the Career Development Chair with the Technion, the Murieland David Jacknow Award for Excellence in Teaching, and the Technions Award for Excellence in Teaching (two times) and several best paper awards and best demo awards together with her research students and colleagues including the SIAM outstanding Paper Prize, the UFFC Outstanding Paper Award, the Signal Processing Society Best Paper Award and the IET Circuits, Devices and Systems Premium Award, and was selected as one of the 50 most influential women in Israel and in Asia. She is also a highly cited Researcher. She was a member of the Young Israel Academy of Science and Humanities and the Israel Committee for Higher Education. She is the Editor-in-Chief for Foundations and Trends in Signal Processing, a member of the IEEE Sensor Array and Multichannel Technical Committee and serves on several other IEEE committees. In the past, she was a Signal Processing Society Distinguished Lecturer, a member of the IEEE Signal Processing Theory and Methods and Bio Imaging Signal Processing technical committees, and served as an Associate Editor for IEEE TRANSACTIONS ON SIGNAL PROCESSING, *EURASIP Journal of Signal Processing*, *SIAM Journal on Matrix Analysis and Applications*, and *SIAM Journal on Imaging Sciences*. She was the Co-Chair and Technical Co-Chair of several international conferences and workshops.