Emil Björnson, Yonina C. Eldar, Erik G. Larsson, Angel Lozano, and H. Vincent Poor

Twenty-Five Years of Signal Processing Advances for Multiantenna Communications

From theory to mainstream technology

©SHUTTERSTOCK.COM/TRIFF

Wireless communication technology has progressed dramatically over the past 25 years, in terms of societal adoption as well as technical sophistication. In 1998, mobile phones were still in the process of becoming compact and affordable devices that could be widely utilized in both developed and developing countries. There were "only" 300 million mobile subscribers in the world [1]. Cellular networks were among the first privatized telecommunication markets, and competition turned the devices into fashion accessories with attractive designs that could be individualized. The service was circumscribed to telephony and text messaging, but it was groundbreaking in that, for the first time, telecommunication was between people rather than locations.

There are now more than six billion subscribers worldwide, and the mobile phone remains the main wireless device, but much has changed. Traditional feature phones with physical keypads have been replaced by smartphones with large touchscreens. Telephony today constitutes a negligible fraction of the traffic, the vast majority of which amounts to packets bearing data for end-user applications. Video and audio streaming, social media, gaming, and a host of other apps, generate the bulk of the traffic. New services continue to arise and cement the smartphone's central role in nearly every aspect of our lives. In parallel, nonhuman-operated devices are progressively coming online to form the Internet-of-Things (IoT) as society continues to be digitized.

Wireless networks have changed dramatically over the past few decades, enabling this revolution in service provisioning and making it possible to accommodate the ensuing dramatic growth in traffic. There are many contributing components, including new air interfaces for faster transmission, channel coding for enhanced reliability, improved source compression to remove redundancies, and leaner protocols to reduce overheads. Signal processing is at the core of these improvements, but nowhere has it played a bigger role than in the development of multiantenna communication. This article tells the story of how major signal processing advances have transformed the early multiantenna concepts into mainstream technology over

Digital Object Identifier 10.1109/MSP.2023.3261505
Date of current version: 1 June 2023

the past 25 years. The story therefore begins somewhat arbitrarily in 1998. A broad account of the state-of-the-art signal processing techniques for wireless systems by 1998 can be found in [2], and its contrast with recent textbooks, such as [3], [4], and [5], reveals the dramatic leap forward that has taken place in the interim.

Fundamentals of multiantenna communications

Traditionally, a base station (BS) at a cellular network site featured antenna panels connected to a baseband unit (BBU) that managed the digital signal processing. These panels, in turn, were tall and narrow, containing multiple vertically stacked radiating elements. By emitting the same signal from such elements, constructive superposition was leveraged to create a radiation pattern, vertically narrow and horizontally wide, that covered a swath of ground in a predefined manner. This is illustrated in Figure 1(a), with each panel's coverage region termed a *cell sector*.

At current BSs, the panels have been replaced with antenna arrays having a more symmetric aspect ratio, which results in radiated beams that can be narrow both horizontally and vertically. The signal transmitted from each antenna element is individually controlled by the BBU, which now has far stronger computational capabilities and can alter the physical shape of the produced beam over both time and frequency. Figure 1(b) illustrates such a setup, and how each beam is narrow enough to aim at a particular user. When these arrays are used in propagation environments with multiple widely spaced paths, each radiated signal loses its directional beam shape and is instead fine-tuned to make the paths superimpose coherently on a small region around the intended receiver.

Antenna arrays bring about three main categories of benefits:

1) *Beamforming gain*: The transmit beam is focused on the receiver, whereby a larger fraction of the radiated energy reaches it. Likewise, multiple receive antennas can collect more energy from selected directions, reinforcing the beam at that end with a focus on the transmitter. The overall beamforming gain is proportional to the transmit and receive array sizes.
2) *Spatial diversity*: There are generally multiple paths via which signals travel between the transmitter and receiver, and the ensuing signal replicas can combine destructively. This causes signal fading, which antenna arrays can mitigate by observing multiple fading realizations simultaneously.
3) *Spatial multiplexing*: Multiple signals can be transmitted concurrently on different beams, either to a single user equipped with multiple antennas, or to multiple users, as in Figure 1(b). This provides a traffic multiplier or *multiplexing gain*, provided the interference among the signals can be kept at bay.

Above, and in the sequel, the beamforming gain is taken as the increase in signal power at the receiver, yet a more nuanced description would further include the reduction in interference to and from unintended users [5, Sec. 5.7]. With a careful design, beamforming can strike an optimum balance between increasing signal energy and reducing interference.

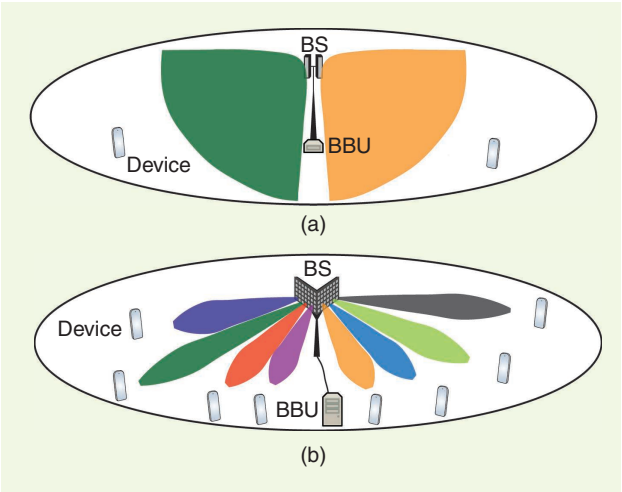

FIGURE 1. (a) A 2G deployment in 1998, consisting of fixed directive antennas that broadcast each signal into a sector. (b) A 5G deployment in 2023, entailing antenna arrays that can exploit the three main multiantenna benefits: beamforming gain, spatial diversity, and spatial multiplexing.

State-of-the-art in 1998

Some of the benefits of antenna arrays were understood well before 1998, but their technology readiness levels were much different than today. Marconi himself famously capitalized on beamforming to enable wireless transatlantic communication in 1901. That experiment relied on an *array antenna*, which achieves beam directivity by connecting multiple elements to the same signal generator. The geometry of the array antenna determines the direction in which the radiated signals superimpose constructively. Hence, the beam direction is fixed and determined at the time of building and erecting the array. This is how the 2G antennas in Figure 1(a) were designed to cover a sector with a fixed beam.

A different beam direction than the one dictated by the array geometry can be realized by emitting the same signal from all of the elements, but with appropriate phase shifts. This concept was first observed experimentally by Ferdinand Braun in 1902, and it led to the phased array technology used for radar since World War II. The phase shifts can be varied over time, to scan for objects in different angular directions. Early field trials of phased arrays for 2G were conducted in 1996 [6]. The possibility of pointing the beams to user locations opened the door to stronger directivities and higher gains, since a beam no longer had to cover an entire sector. The difference between array antennas and phased arrays is illustrated in Figure 2, which also depicts the digital antenna arrays featured in 5G, where each element is connected to a separate signal generator.

In parallel with the refinement of phased arrays for beamforming over several decades, the use of multiple receive antennas for diversity also became commonplace. Spatial diversity was conceived for signal reception as far back as the 1930s [7] and builds on an intuitive principle: if the same signal reaches several physically separated antennas, it is unlikely

that the multipath propagation environment causes destructive superposition, hence signal fading, at all such antennas simultaneously. By decoding a combination of the observations at the various receive antennas, the communication becomes much more reliable.

A wireless network must of course provide an uplink connection from users to BSs, as well as a downlink connection from BSs to users. Thus, diversity is desirable in both link directions, yet transmit diversity did not emerge until the 1990s [8]. In 1998, the Alamouti space–time block code for two antennas was proposed [9] and a more general framework for space–time coding with multiple transmit antennas was published soon thereafter [10]. The principle is to repeatedly transmit a block of data symbols while varying the spatial directivity in a predetermined way (e.g., using different antennas); the receiver collects observations over a time interval and decodes them. Space–time codes are carefully crafted to not only enable decoding, but to strike a satisfactory tradeoff between high spectral efficiency (i.e., bits per second per Hertz of spectrum), high diversity, and low complexity.

Altogether, beamforming gains and spatial diversity were known by 1998, and it was largely thought that these were the two main benefits of antenna arrays: beamforming gains in the case of coherent arrays, associated with tight antenna spacings and cleanly defined directions of arrival and departure, and diversity gains in the case of arrays experiencing largely uncorrelated fading across the antennas, associated with wider spacings and rich multipath settings. The third, and ultimately the most powerful benefit of antenna arrays, spatial multiplexing, was still largely under the radar. However, its seeds had already been planted in research efforts on interference-aware beamforming [11] and on communication concepts for linear channels that couple multiple inputs into multiple outputs [12]. Unlike beamforming and diversity, which involved replicas of a single signal, these precursors of spatial multiplexing entailed the transmission and reception of distinct signals simultaneously and on the same bandwidth. Particularly prescient was the transmission and reception with two orthogonally polarized antennas, subsequently extended in a piece that featured multiantenna transmitters and receivers with many of the ingredients required for true spatial multiplexing [13]. However, it was not until after 1998 that all of these pieces fell into place.

Wireless networks have changed dramatically over the past few decades, enabling this revolution in service provisioning and making it possible to accommodate the ensuing dramatic growth in traffic.

External technology developments

Three external trends have heavily guided and influenced the evolution of multiantenna technology over the past decades.

1) The explosion in wireless traffic, which has doubled every 18 months as per Cooper's law, along with a fundamental change in the nature of such traffic, was driven by new user behaviors and applications. An efficient network for telephony had to support many simultaneous fixed-rate connections, while today's data networks aim at maximizing the bit rate per user device (to support certain applications) and the bit rate per unit area (to accommodate many devices).
2) The exponential improvement and size reduction of integrated circuits have led to systems-on-a-chip that combine radios, memory, and processors capable of advanced

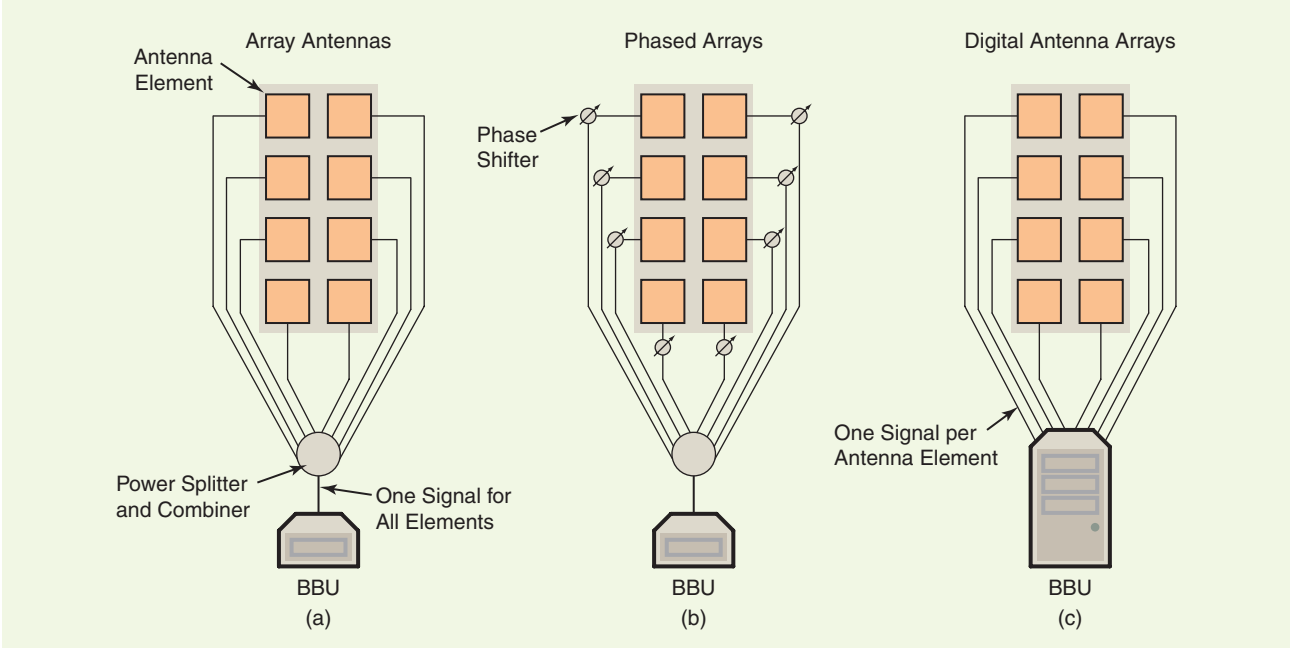

FIGURE 2. There are three classical categories of arrays: (a) array antennas that generate fixed beams, (b) phased arrays that rely on phase shifters to control the beam direction, and (c) digital antenna arrays that have full control of the signal transmitted from each antenna element.

signal processing on a tiny piece of silicon. While in 1998, a digital antenna array with $M = 2$ or $M = 4$ elements would consist of M external antenna elements connected to M radio frequency (RF) units and one BBU, current 5G BSs can integrate $M = 64$ elements and RF units into a single box. This development has also enabled smartphones to feature digital arrays, for now with $M = 4$ elements.

3) The gradual change in the signal waveforms: There were multiple 2G standards based either on time-division multiple access (TDMA) or code-DMA (CDMA). The first versions of 3G, finalized precisely around 1998, were entirely based on CDMA, which won the battle against the competing orthogonal frequency-DMA (OFDMA). For 4G, the shift to OFDMA finally took place, and 5G retained this same waveform after a handful of alternatives were evaluated and discarded. While all waveforms are in principle compatible with antenna arrays, the choice does have a fundamental impact on what signal processing algorithms are required.

Marconi himself famously capitalized on beamforming to enable wireless transatlantic communication in 1901.

Five key areas of signal processing advances

We have identified five stages of signal processing advances in the evolution of multiantenna technology from 2G to 5G and beyond. The background, new solutions, and specific insights are expounded on in the following sections.

From spatial diversity to spatial multiplexing

One could argue that, all the way back to Marconi, beamforming was motivated by the interest in extending the range of coverage. In turn, diversity was motivated by the desire to increase reliability. By 1998, the exploding cost of radio spectrum ahead of 3G brought about a new and powerful necessity: increasing the spectral efficiency. The shift toward high bit-rate user applications further amplified this trend. The operational mode of antenna arrays that maximizes the spectral efficiency is spatial multiplexing and, after 1998, the atmosphere was therefore primed for it to finally come to the fore.

The prerequisite for spatial multiplexing is a multiple-input multiple-output (MIMO) communication channel, where each input/output refers to an antenna element in a digital antenna array. There are two MIMO categories: *single-user MIMO* entails a multiantenna BS and a multiantenna user device, while *multiuser MIMO* encompasses a multiantenna BS and multiple user devices.

Arguably, the main catalyst for single-user MIMO was the work in [14], which set out to design the perfect transceivers from an information-theoretic standpoint. Starting with transmit and receive digital antenna arrays and no preset conditions on how to employ them, it was found that, if the elements within each array exhibited uncorrelated fading, the optimum strategy was to have each radiate an independent data-carrying signal. This was radically novel in that it sought to exploit, rather than counter, multipath propagation; it is the very existence of multiple paths that allows the receiver to observe a distinct linear combination of the transmit signals at each receive antenna, from where those transmit signals can be resolved. The number of signals that can be spatially multiplexed is then limited by the minimum of the number of transmit and receive antenna elements. In follow-up work, a specific architecture was proposed to effect such spatial multiplexing, the so called layered architecture, which was remarkable in that it could be built with off-the-shelf encoders and decoders and did not require the transmitter to know anything about the channel [15]. Additional results progressively solidified the theoretical underpinnings [16]. In particular, the idea of transmitting concurrent signals, one from each antenna element, was generalized to the transmission of concurrent beams from all elements at once. Phased arrays cannot achieve such spatial multiplexing because they only create one beam at a time, and digital antenna arrays are decidedly necessary.

Multiuser MIMO can be traced back to signal processing concepts for simultaneous uplink reception from multiple users [17] and simultaneous downlink beamforming to users in different angular directions [18]. Here, the number of signals that can be spatially multiplexed is not limited by the number of antenna elements per user, but rather across all users; even if each user features a single element, it is possible to spatially multiplex one signal to/from each one. This major advantage comes at the expense of the BS having to carefully arrange the transmit and receive beams, such that each one matches with the multipath characteristics of its intended user and there is minimal interference among them, as in Figure 2(b). With that, every user can transmit continuously and over the entire system bandwidth, rather than only in a time slot and/or frequency subband, reflecting the spatial multiplexing benefit. Multiuser MIMO is a generalization to multipath settings of classical space-DMA (SDMA), whereby users share a channel in space rather than in time or frequency. Interestingly, the SDMA concept is more than 20 years older than single-user MIMO [19], which showcases that establishing many simultaneous user connections was long perceived more important in wireless networks than achieving high data rate per connection.

The potential of MIMO, in both its single-user and multiuser fashions, sparked a chain reaction that spread rapidly through academia and industry, bringing much excitement by the early 2000s. Cellular standardization bodies, in particular, the 3G partnership project (3GPP) adopted it in a limited fashion for late 3G releases and then as an integral part of the designs beginning with 4G. Even faster was the adoption within Wi-Fi, with the first version including MIMO certified in 2007 and supported by a multitude of devices, including laptops, tablets, and smartphones.

MIMO harnesses the three dimensions of benefits shown in Figure 3: beamforming gain, spatial diversity, and spatial multiplexing. A clear understanding of how these benefits are

related has emerged over time. Fundamental tradeoffs have been identified, for given array configurations and channel conditions.

- *Beamforming* is a special case of spatial multiplexing where a single beam is transmitted to a single user. This is in fact the optimum strategy when the signal-to-noise ratio (SNR) is low; maximizing the signal energy is then of the essence, and the best recipe is to concentrate all the radiated energy on the strongest beam. At high SNR, in contrast, energy is plentiful and can be spread over multiple beams, to the point that it is optimum to activate as many beams as the channel and antenna counts allow. This is represented by the blue plane in Figure 3.
- *Spatial diversity*, roughly quantified as the number of independently faded signal replicas, and *spatial multiplexing*, meaning the number of concurrent beams, cannot be simultaneously maximized. These two quantities are rather subject to a tradeoff [20]. At the extreme points of this diversity-multiplexing tradeoff, one of the quantities is maximized while the other stands at a minimum. Various combinations are feasible at intermediate points of the tradeoff curve, which is cartooned on the yellow plane of Figure 3.

While all waveforms are in principle compatible with antenna arrays, the choice does have a fundamental impact on what signal processing algorithms are required.

As mentioned, the choice between beamforming and spatial multiplexing is dictated by the SNR, hence, by its underlying parameters (e.g., transmit power, channel attenuation, noise power). In turn, the mix of diversity and multiplexing depends on whether the priority is to increase reliability or spectral efficiency. However, the operating point should be selected holistically, and over the years this has caused the mix to shift toward less diversity and more multiplexing. Indeed, as successive system generations have spanned ever broader bandwidths, more and more diversity has been reaped in the frequency domain. The rewards of additional diversity rapidly saturate, thus the need for spatial diversity has abated [21]. This of course does not apply to narrowband control channels or to low-power short-packet IoT communication, where spatial diversity remains important, but it does hold for the user data channels that carry the bulk of the traffic in cellular networks.

From spatial multiplexing to massive MIMO

The basics of MIMO, developed under the premises of perfect channel state information (CSI) and rich scattering, indicate that an arbitrarily high spectral efficiency can be achieved by deploying sufficiently many antennas and serving many users at once. However, the practical challenges became apparent when the technology was first commercialized. The spatial multiplexing capability in single-user MIMO was often restricted by limited scattering, while multiuser MIMO is restrained by imperfect CSI. Massive MIMO, a new form of multiuser MIMO that originated from [22], was developed in the 2010s to address these issues and is now at the heart of 5G.

The new aspects of massive MIMO are as follows. First, it relies on having many more BS antennas than spatially multiplexed users. This design choice renders the beams relatively narrow (e.g., in the sense of focusing on a small region around the intended receiver), hence there is likely to be little overlap among beams focused on distinct users. Moreover, by virtue of these favorable conditions, whatever little interference exists can be suppressed through low-complexity linear signal processing: for example, regularized zero-forcing that fine-tunes each beam's focal area to balance a strong beamforming gain with low interference [23]. At the same time, and again because of the excess BS antennas, the effective channels provided by these beams harden, meaning that they become very stable and subject to only minimal fading fluctuations.

Second, massive MIMO is tailored for resource-efficient CSI acquisition. The main estimation principle is to emit separate predefined pilot signals from each antenna element and then gauge the channel coefficients from the observations of these pilots at the receive elements. Massive MIMO adopted time-division duplexing (TDD), where the same bandwidth is utilized, in alternating fashion, for uplink and downlink. Since, by virtue of reciprocity, the channel is then identical in both directions, it suffices to estimate its coefficients in one direction. Specifically, the CSI required for both uplink and downlink is obtained from uplink pilots. The necessary pilot resources are thus determined by the number of multiplexed users, with no dependence on the number of BS antennas. In contrast, many previous commercial implementations of multiuser MIMO were based on frequency-division duplexing (FDD), where the uplink and downlink channels were

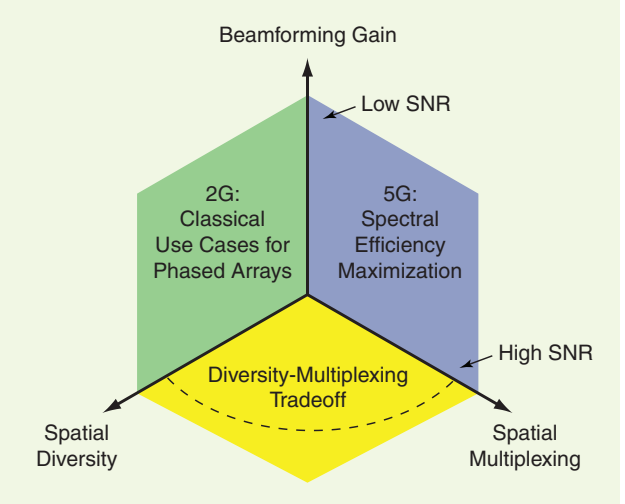

FIGURE 3. There are three benefit dimensions of multiantenna communication that have developed in past decades. From 2G to 5G, systems have shifted from the green plane to the blue plane; that is, spatial diversity has gradually been replaced by spatial multiplexing. Since spatial multiplexing requires high SNRs to be practically useful, beamforming gains remain essential at low SNRs.

entirely different, or TDD operation without using reciprocity. The downlink operation then required the BS to transmit as many pilots as it has antennas, except in specific propagation scenarios where the channels can be parametrized using a few angles. Moreover, each user needed to quantize and feedback its channel estimates to the BS. In a typical 5G setup with $M = 64$ BS antennas that spatially multiplex $K = 8$ users, the FDD alternative would require $M/K = 8$ times as many pilots and a proportional amount of extra CSI feedback.

The TDD operation is particularly helpful in complex propagation environments with many paths per user, such as the one sketched in Figure 4, where the optimum downlink transmission spreads a user's signal energy in many directions to match the reflecting objects. CSI acquisition through uplink pilots automatically captures these fine characteristics, without any prior channel knowledge or array calibration. In FDD operation, besides requiring vastly more pilot resources, essential channel details are lost in the feedback quantization. In cellular networks, pilot signals must be reused with care across cells to avoid pilot contamination phenomena, whereby BSs inadvertently beamform toward pilot-sharing users in neighboring cells. This is particularly a concern in TDD operation, where uplink estimation errors also affect the downlink. A multitude of signal processing and resource allocation schemes have been developed over the past decade to alleviate pilot contamination [3], [4].

Massive MIMO provides a solid foundation for practical signal processing design. While sophisticated information theory for multiuser MIMO existed already by the 2000s [24], it was largely limited to scenarios with perfect CSI. As CSI quality is the main limiting factor of multiuser MIMO performance, this constrained the practical usefulness of the available theory. Thanks to the reliance on linear signal processing, massive MIMO analyses successfully handle imperfect CSI and hardware imperfections, resulting in rigorous and mathematically clean spectral efficiency expressions that not only predict actual performance accurately, but serve as effective tools for system optimization (e.g., pilot allocation, power control, and beamforming). Massive MIMO theory not only turned multiuser MIMO into a practically feasible technology, the analytical elegance also expanded the way information theory for wireless communication can be taught [3], [4].

The potential of MIMO, in both its single-user and multiuser fashions, sparked a chain reaction that spread rapidly through academia and industry, bringing much excitement by the early 2000s.

As mentioned, uplink–downlink reciprocity in TDD operation is important for massive MIMO. Reciprocity holds for the over-the-air propagation as long as the channel impulse response remains constant: that is, provided the duplexing takes place within the channel coherence time. However, the transceiver hardware is generally not reciprocal between transmission and reception, for instance due to mismatches in the local oscillators. Such hardware nonreciprocity calls for a calibration procedure that phase-synchronizes the antennas within each array through occasional mutual measurements.

Today, 5G BSs feature almost exclusively massive MIMO configurations in TDD bands, with arrays of $M = 32$ or $M = 64$ antennas being the most common. Early on, there were concerns that the signal processing would entail an exceedingly high energy consumption, but this concern was later dispelled, and dedicated systems-on-chip are now available that implement clever signal processing algorithms for the entire BBU, including massive MIMO, at reasonable energy costs.

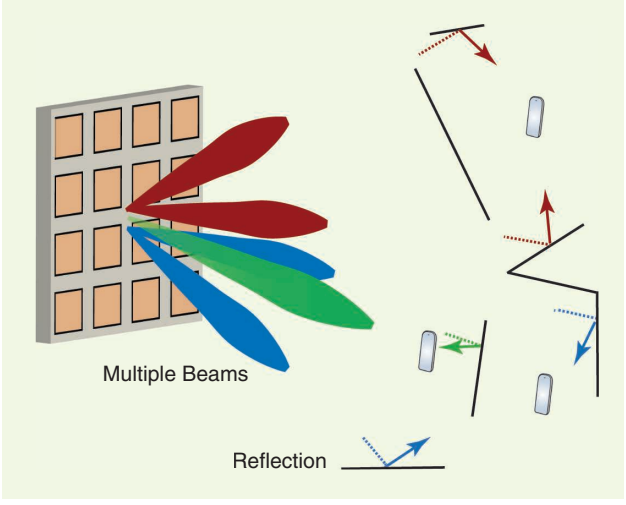

FIGURE 4. The propagation channel consists of a multitude of specularly or diffusely reflecting objects. Such channels are very challenging to estimate with sufficient accuracy to enable multiuser MIMO communication, but the TDD CSI estimation approach in massive MIMO manages this.

A quest for more bandwidth at higher frequencies

Bit rate has long been the performance metric that users of wireless technology are most familiar with, and hence wireless technology has evolved to support higher values thereof. The bit rate enjoyed by a single device equals the product of the spectral efficiency and the spectral bandwidth. Therefore, besides being driven higher by single-user MIMO, bit rates have expanded over time thanks to the allocation of new frequency bands. A 2G network typically had access to 20 MHz at carrier frequencies around 1 GHz, while current 5G networks primarily span 100 MHz in the 3.5-GHz range, with the standard supporting in excess of 500 MHz.

The radio spectrum is a limited natural resource shared by a multitude of technologies, including those beyond the civilian wireless communication arena considered in this article. While a few sub-6-GHz bands have been refarmed from outdated technologies to cellular networks, the strive for fresh bandwidth inevitably pushes systems toward ever higher frequencies. In particular, millimeter-wave (mmWave) bands, nominally starting at 30 GHz, are now part of 5G. First-generation mmWave technology has been rolled out by a few telecom

operators, while the bulk of them wait for the hardware to mature and for the 3.5-GHz band to become congested.

The field strength of a signal radiated from a point source in free space attenuates with distance in a frequency-independent manner. Then, the power captured from such an electric field is proportional to the receiver's aperture and, since the size of an antenna element shrinks with the wavelength, an increased carrier frequency necessitates further antenna elements to maintain the desired aperture. Moreover, the channel conditions in cellular networks become steadily more challenging as the frequency shifts up due to reduced scattering and diffraction, and steeper penetration losses, all of which call for beamforming gains. Multiantenna technology is therefore paramount at mmWave frequencies.

At the same time, implementation becomes difficult, and not only because of the added hardware components and the huge dimensionalities in digital signal processing. When moving to higher frequencies and broader bandwidths, hence to faster sampling rates, power amplifier efficiency and dissipation in analog-to-digital converters (ADCs) are further issues that need attention. The signal processing community has explored two main ways to deal with the hardware and algorithmic complexity [25].

The first option is to reduce the number of RF units, particularly converters, by designing transceivers as a mix of phased arrays and digital antenna arrays. The resulting hybrid analog–digital antenna array is illustrated in Figure 5, where each column is a phased subarray that is connected separately to the BBU, such that different signals can be transmitted and received. Each subarray can form a single beam and spatial multiplexing can then be applied through different linear combinations of those beams. If the channel features a small number of propagation paths, each subarray can focus a beam on one of those paths, and the communication performance of a digital antenna array can be attained with fewer hardware components. In the multipath scenario illustrated in Figure 4, five distinct beam directions are sufficient to communicate effectively. Hybrid antenna arrays can take other forms, but generally entail a semianalog beamforming implementation with more antenna elements than digital ports [26]. There are several prices to pay for abandoning the digital antenna array paradigm. One can only transmit as many beams as there are digital ports and the beamforming fidelity is crippled in wideband systems since combinations of the same beams must be used on all subcarriers. Channel estimation becomes more intricate since each phased array must sweep through as many beam directions as it has elements in order to excite all channel dimensions.

An alternative to reducing the number of RF units is digital antenna arrays with lowered ADC resolutions, as also illustrated in Figure 5. The energy consumption of an ADC grows exponentially with the resolution, hence enormous energy reductions are possible by moving from the conventional 15 bits per sample down to, say, five bits per sample. And yet, since an array with M elements and b-bit ADCs collects a total of bM bits per sample period, the total number of ADC bits can still be sizeable even if b is small, explaining why a high spectral efficiency can be maintained [27]. The extreme case of uplink massive MIMO with $b = 1$ happens to be analytically tractable [28], which has facilitated the emergence of signal

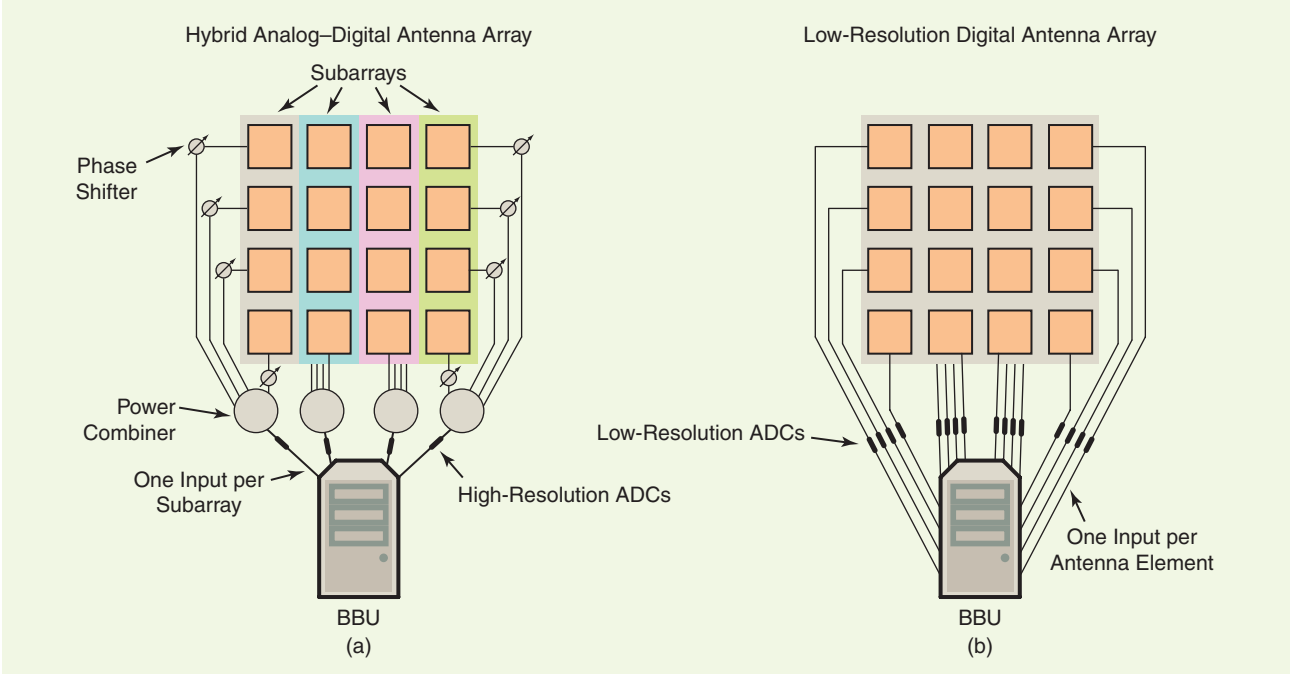

FIGURE 5. Conventional digital antenna arrays incur a high energy consumption when implemented at mmWave frequencies. There are two main ways to circumvent this: (a) Reduce the number of components using hybrid analog-digital antenna arrays, consisting of multiple phased subarrays; (b) Design digital antenna arrays with simplified components, such as low-resolution ADCs.

processing algorithms that compensate for the ensuing quantization distortion. The downlink counterpart involves low-resolution digital-to-analog converters [29].

Reduced bit resolution is also viable for hybrid antenna arrays; if the analog signal combining prior to quantization and the digital postprocessing is properly optimized for the communication task at hand, the signal content becomes more amendable to a low-bit representation [30].

The MIMO technology is now present in every smartphone and BS that enters the market.

The initial 5G mmWave products are based on hybrid arrays, but there are indications that the low-resolution approach might eventually become the preferred solution [31].

Further opportunities for dimensionality reduction

The joint evolution toward arrays with more antenna elements and wider bandwidths, requiring higher sampling rates, makes it essential not to overdesign the transceivers. As an alternative to scaling up a conventionally small digital array, Figure 5 showcases two ways of increasing the antenna element counts while reducing the hardware complexity per element. Both approaches capitalize on the massive MIMO philosophy of having more antenna elements than multiplexed beams, which enables a reduction in the beamforming exactness because the beams are so narrow that interuser interference is low anyway. Further dimensionality reductions are possible by exploiting the structure of the channel—say, sparsity in the angular, frequency, and time domains—or by exploiting the specific task the system is designed to address.

Besides being exponential in the resolution, the energy consumption of an ADC is proportional to the sampling rate. A host of ideas based on sub-Nyquist sampling and compressed sensing have been proposed to reduce the sampling rate by exploiting various forms of channel structure that exist in many scenarios [32]. In mmWave channels with a small number N of propagation paths, there might be only roughly N nonzero taps in the channel impulse response regardless of the bandwidth. Such time-domain sparsity in the channel response can be leveraged to reduce the sampling rate [33]. The N paths are likely distinct also in the angular domain, given that BSs are usually deployed high above the environmental clutter; reflections take place only locally around each user, subtending a small angular spread at the distant BS. Figure 6 illustrates such a scenario with $N = 3$ distinct paths, supporting three multiplexed signals. The line-of-sight path has a distinctly short delay, while the two remaining paths have similar delays but are clearly distinguishable in the angular realm. The joint channel sparsity is represented by the three colored entries in a time-angle matrix, with the vast majority of entries containing no propagation paths. Combining these forms of sparsity with modern compressed-sensing tools enables hefty reductions in sampling rates.

Many data services exhibit intermittent activity patterns; among the thousands of devices associated with a BS, only a small subset requires data transfers within a given time slot. Signal processing can enable these devices to transmit efficiently without requiring a preceding access procedure. The key is to assign each device with a unique but nonorthogonal pilot sequence and then utilize sparsity in the user domain along and the large number of spatial samples obtained over an antenna array to enable user identification and channel estimation [34]. The joint user and data detection problem has also been approached using compressed sensing methods [35].

Commercial massive MIMO products already exploit some elementary channel sparsity; for instance, the received signals over the many antennas might be transformed into an equal number of angular dimensions. The dimensions that contain little power are discarded in the early stages of the digital

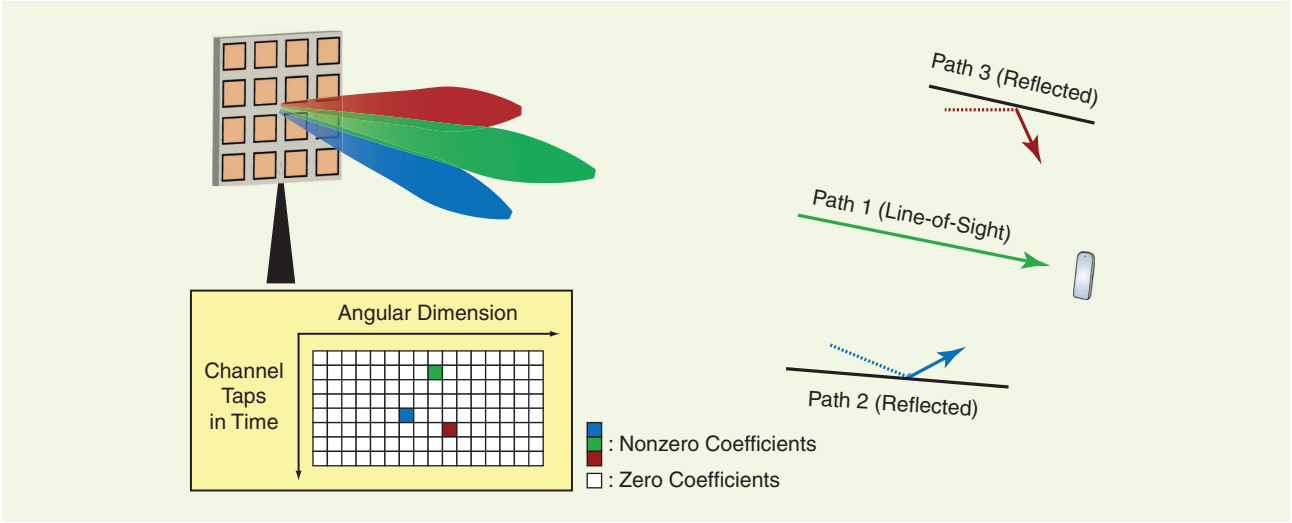

FIGURE 6. The channel in mmWave systems with large spectral bandwidths and antenna counts might exhibit sparsity in the joint time-angle domain. In this example, there are $N = 3$ paths that are distinct in both time and angle. The sparse impulse response can be exploited along with compressed sensing techniques to reduce the sampling rate, thereby lowering power consumption.

uplink processing to shrink the dimensionality of the remaining computations. However, the more radical compressed sensing solutions are yet to be brought to life.

Machine learning-based algorithmic refinements

One of the most active areas in contemporary signal processing is machine learning (ML). While this is a many decades-old discipline, the increased availability of large amounts of data and processing power has, in recent years, greatly enhanced its potential to transform the implementation of many signal processing tasks from more traditional model-driven algorithms into data-driven ones. This transformation is also taking place in the context of signal processing for wireless communications, which has traditionally been very heavily (and successfully) model-based. Several trends are driving this transformation. One trend is that, with the vast amount of IoT and machine-type connections that coexist with human-type broadband connections, wireless network traffic is becoming increasingly intricate to model accurately, thereby making network operation difficult to optimize. Another trend is that antenna arrays and other sensors are becoming pervasive on smartphones and other connected devices, hence the volume of data available for learning is swelling dramatically. Yet a third trend is that the amount of processing power distributed throughout wireless networks is growing rapidly, giving rise to paradigms such as fog and edge computing.

Recent theoretical breakthroughs, including ML-based algorithms, are bound to continue sustaining the progress of the technology.

There is a confluence of ML and communications in the optimization of wireless networks. This is a very natural application for ML since the operation of these networks involves a multitude of tasks that ML is good at addressing, including inferential tasks, such as channel estimation, signal detection, and data decoding, as well as decision-making tasks such as routing, access control, and resource allocation. ML-based solutions can capture practical characteristics that were overlooked by the models, underpinning existing algorithms. However, to ensure that ML algorithms improve upon the existing, it is essential to initiate the training procedure judiciously.

The model-aided ML paradigm provides a structured way to transfer classical know-how from the signal processing community onto new ML algorithms [36]. Figure 7 exemplifies how an existing iterative algorithm can be transformed into an enhanced ML algorithm. The existing algorithm takes an initial input signal and processes/updates it iteratively until a predefined termination criterion is satisfied, at which point the final output is obtained. The specific processing is normally obtained through model-based algorithm design. Instead of expressing the algorithm as a loop, L iterations of the algorithm can be expressed as a sequence of L identical processing layers. If a data-driven training procedure is employed to fine-tune these processing layers, which no longer have to be identical, what ensues is an ML algorithm that is guaranteed to perform better than the original model-based algorithm. This procedure is called *algorithm unrolling* or *deep unfolding*, and it has in recent years been utilized to enhance various multiantenna tasks, including signal detection [37] and beamforming optimization for downlink multiuser MIMO [38].

A peek into the future

Over the past 25 years, multiantenna techniques have gone from rudimentary designs for beamforming and diversity combining to a mainstream technology that uses massive

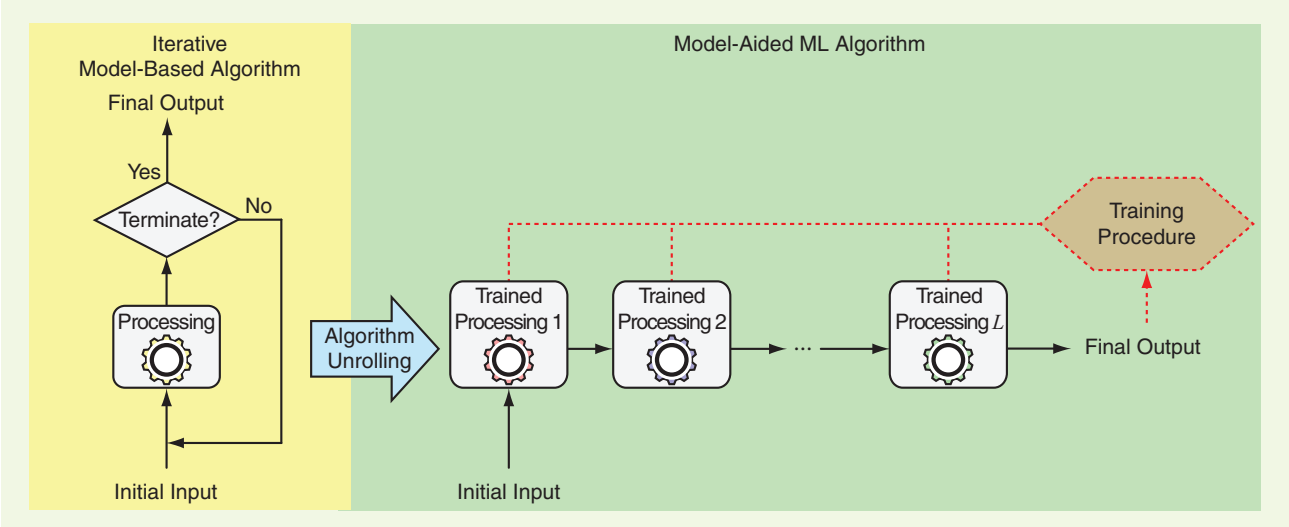

FIGURE 7. Many conventional model-based algorithms for optimization in multiantenna communication are iterative. Suppose one such algorithm can be expressed as an iterative processing loop that continues until a termination criterion is satisfied. An enhanced ML algorithm can be developed through algorithm unrolling; that is, writing the iteration as L separate processing layers and fine-tuning these layers through a data-aided training procedure.

spatial multiplexing to multiply the capacity of 5G networks. The MIMO technology is now present in every smartphone and BS that enters the market. Fast and capable signal processing algorithms have enabled this leap forward, and are currently buttressing the emergence of low-power 5G mmWave transceivers where high-resolution hardware components are replaced with digital processing. Recent theoretical breakthroughs, including ML-based algorithms, are bound to continue sustaining the progress of the technology.

Indeed, we expect the multiantenna communication journey to continue. The insatiable growth in data traffic can only be met by deploying ever more antennas and ever more bandwidth. The massive MIMO philosophy prescribes that the number of BS antennas, M, must scale proportionally to the number of active users, K. As the complexity of algorithms, such as regularized zero-forcing is proportional to MK^2 [4], a linear scaling in both K and M implies a cubic complexity growth. Moreover, once the bandwidth surpasses 1 GHz, the sampling rates approach the clock speed of existing processors, which renders the implementation even more demanding. The compressed sensing algorithms described earlier might be suitable to address these challenges, but there is likely room for many new signal processing advances.

When adding ever more antennas to BSs, practical size and weight constraints might make new deployment principles necessary, beyond the boxes-in-a-tower paradigm. One promising approach is to distribute the antennas over multiple physical locations while retaining the coherent transmission and reception processing, a concept rooted in cell cooperation and network MIMO ideas, as well as in the notion of remote RF units, and whose present embodiment is termed *cell-free massive MIMO* [39]. Apart from stronger beamforming gains, a distributed antenna deployment can provide improved spatial multiplexing capabilities and macroscopic diversity against the shadowing of large objects in the environment. The current trend of shifting baseband computations from BS sites to edge-cloud computers will ease the adoption of this deployment approach.

After two decades of smartphones ruling the wireless ecosystem, other devices, such as extended reality eyeglasses, are predicted to take center stage. New services will surface, with renewed standards for the bit rates, latency, and reliability that users expect wireless networks to deliver. Other performance metrics might arise to dictate future technology development, particularly related to sustainability, environmental impact, and deployment costs, as well as to the digital divide between the digitized and far-from-digitized regions of the world.

Beyond the signal processing advances captured in this article, two emerging research topics build on multiantenna technology. The first is integrated communication and sensing [40], which explores how large-scale antenna arrays can be simultaneously used for accurate radar sensing, localization, and communication. It seems natural that the deployment of massive antenna numbers for communication purposes can be the catalyst for other applications that benefit from wireless measurements. Another related research direction is that of smart surfaces [41], whose signal reflection properties can be controlled by means of metamaterials with programmable impedance patterns. These reconfigurable intelligent surfaces provide a sort of passive beamforming that is particularly useful to enhance propagation conditions over wireless channels.

Authors

Emil Björnson (emilbjo@kth.se) is a professor of wireless communication at the KTH Royal Institute of Technology, Stockholm, Sweden. He has authored three textbooks on multiple-input multiple-output technology, has received 23,000 citations, and has published a large amount of simulation code. He has received the 2018 and 2022 IEEE Marconi Prize Paper Awards in Wireless Communications, the 2019 EURASIP Early Career Award, the 2019 IEEE Communications Society Fred W. Ellersick Prize, the 2019 IEEE Signal Processing Magazine Best Column Award, the 2020 Pierre-Simon Laplace Early Career Technical Achievement Award, the 2020 CTTC Early Achievement Award, and the 2021 IEEE ComSoc RCC Early Achievement Award. He is an IEEE Fellow.

Yonina C. Eldar (yonina.eldar@weizmann.ac.il) is a professor in the Department of Math and Computer Science at the Weizmann Institute of Science, Rehovot, Israel, where she heads the Center for Biomedical Engineering and Signal Processing. She is also a visiting professor at the Massachusettes Institute of Technology and at the Broad Institute, Cambridge, MA 02142 USA, and an adjunct professor at Duke University, Durham, NC 27708 USA, and was a visiting professor at Stanford University, Stanford, CA USA. She is a member of the Israel Academy of Sciences and Humanities and a European Association for Signal Processing Fellow. She has received many awards for excellence in research and teaching and heads the Committee for Promoting Gender Fairness in Higher Education Institutions in Israel. She is an IEEE Fellow.

Erik G. Larsson (erik.g.larsson@liu.se) is a professor at Linköping University, Linköping, Sweden. He coauthored the textbook *Fundamentals of Massive MIMO* (Cambridge University Press, 2016). He received, among others, the IEEE ComSoc Stephen O. Rice Prize in Communications Theory in 2015, the IEEE ComSoc Leonard G. Abraham Prize in 2017, the IEEE ComSoc Best Tutorial Paper Award in 2018, and the IEEE ComSoc Fred W. Ellersick Prize in 2019. His interest include wireless communications, statistical signal processing, and networks. He is an IEEE Fellow.

Angel Lozano (angel.lozano@upf.edu) received his Ph.D. from Stanford University in 1999, worked for Bell Labs (Lucent Technologies, now Nokia) between 1999 and 2008, and served as an adjunct associate professor at Columbia University between 2005 and 2008. He is a professor at Universitat Pompeu Fabra, Barcelona, Spain. His papers have received several awards, including the 2009 Stephen O. Rice Prize, the 2016 Fred W. Ellersick Prize, and the 2016 Communications Society & Information Theory Society Joint Paper Award. He is also the recipient of a European Research

Council Advanced Grant for the period 2016–2021 and a 2017 Highly Cited Author. He is the coauthor of the textbook *Foundations of MIMO Communication*, published by Cambridge University Press in 2019. He is an IEEE Fellow.

H. Vincent Poor (poor@princeton.edu) is the Michael Henry Strater University Professor at Princeton University, Princeton, NJ USA, where his interests include information theory, machine learning, and network science, and their applications in wireless networks, energy systems, and related fields. Among his publications in these areas is the recent book *Machine Learning and Wireless Communications* (Cambridge University Press, 2022). He is a member of the U.S. National Academies of Engineering and Sciences, and received the IEEE Alexander Graham Bell Medal in 2017. He is an IEEE Life Fellow.

References

[1] "World telecommunication development report: Mobile cellular," International Telecommunication Union, Geneva, Switzerland, Tech. Rep. 5, 1999.

[2] H. V. Poor and G. W. Wornell, *Wireless Communications: Signal Processing Perspectives* (Prentice-Hall Signal Processing Series). Upper Saddle River, NJ, USA: Prentice-Hall, 1998.

[3] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, *Fundamentals of Massive MIMO*. Cambridge, U.K.: Cambridge Univ. Press, 2016.

[4] E. Björnson, J. Hoydis, and L. Sanguinetti, "Massive MIMO networks: Spectral, energy, and hardware efficiency," *Found. Trends® Signal Process.*, vol. 11, nos. 3–4, pp. 154–655, Nov. 2017, doi: 10.1561/2000000093.

[5] R. W. Heath Jr. and A. Lozano, *Foundations of MIMO Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2018.

[6] S. Anderson, U. Forssen, J. Karlsson, T. Witzschel, P. Fischer, and A. Krug, "Ericsson/Mannesmann GSM field-trials with adaptive antennas," in *Proc. IEE Colloq. Adv. TDMA Techn. Appl.*, 1996, pp. 1–6, doi: 10.1049/ic:19961235.

[7] H. O. Peterson, H. H. Beverage, and J. B. Moore, "Diversity telephone receiving system of R.C.A. communications, Inc.," in *Proc. Inst. Radio Eng. (IRE)*, 1931, vol. 19, no. 4, pp. 562–584, doi: 10.1109/JRPROC.1931.222363.

[8] A. Wittneben, "Basestation modulation diversity for digital simulcast," in *Proc. 41st IEEE Veh. Technol. Conf. (VTC)*, 1991, pp. 848–853, doi: 10.1109/VETEC.1991.140615.

[9] S. Alamouti, "A simple transmit diversity technique for wireless communications," *IEEE J. Sel. Areas Commun.*, vol. 16, no. 8, pp. 1451–1458, Oct. 1998, doi: 10.1109/49.730453.

[10] V. Tarokh, N. Seshadri, and A. Calderbank, "Space-time codes for high data rate wireless communication: Performance criterion and code construction," *IEEE Trans. Inf. Theory*, vol. 44, no. 2, pp. 744–765, Mar. 1998, doi: 10.1109/18.661517.

[11] J. H. Winters, "Optimum combining in digital mobile radio with cochannel interference," *IEEE J. Sel. Areas Commun.*, vol. 2, no. 4, pp. 528–539, Jul. 1984, doi: 10.1109/JSAC.1984.1146095.

[12] J. Salz, "Digital transmission over cross-coupled linear channels," *AT&T Tech. J.*, vol. 64, no. 6, pp. 1147–1159, Jul./Aug. 1985, doi: 10.1002/j.1538-7305.1985.tb00269.x.

[13] J. Winters, "On the capacity of radio communication systems with diversity in a Rayleigh fading environment," *IEEE J. Sel. Areas Commun.*, vol. 5, no. 5, pp. 871–878, Jun. 1987, doi: 10.1109/JSAC.1987.1146600.

[14] G. J. Foschini and M. J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," *Wireless Pers. Commun.*, vol. 6, no. 3, pp. 311–335, Mar. 1998, doi: 10.1023/A:1008889222784.

[15] P. W. Wolniansky, G. J. Foschini, G. D. Golden, and R. A. Valenzuela, "V-BLAST: An architecture for realizing very high data rates over the rich-scattering wireless channel," *in Proc. URSI Int. Symp. Signals, Syst., Electron. Conf. Proc. (Cat. No.98EX167)*, Sep. 1998, pp. 295–300, doi: 10.1109/ISSSE.1998.738086.

[16] E. Telatar, "Capacity of multi-antenna Gaussian channels," *Eur. Trans. Telecommun.*, vol. 10, no. 6, pp. 585–595, Nov./Dec. 1999, doi: 10.1002/ett.4460100604.

[17] J. H. Winters, "Optimum combining for indoor radio systems with multiple users," *IEEE Trans. Commun.*, vol. 35, no. 11, pp. 1222–1230, Nov. 1987, doi: 10.1109/TCOM.1987.1096697.

[18] S. C. Swales, M. A. Beach, D. J. Edwards, and J. P. McGeehan, "The performance enhancement of multibeam adaptive base-station antennas for cellular land mobile radio systems," *IEEE Trans. Veh. Technol.*, vol. 39, no. 1, pp. 56–67, Feb. 1990, doi: 10.1109/25.54956.

[19] Y. Tsuji and Y. Tada, "Transmit phase control system of synchronization burst for SDMA/TDMA satellite communication system," U.S. Patent 3 995 111, 1976.

[20] L. Zheng and D. Tse, "Diversity and multiplexing: A fundamental tradeoff in multiple-antenna channels," *IEEE Trans. Inf. Theory*, vol. 49, no. 5, pp. 1073–1096, May 2003, doi: 10.1109/TIT.2003.810646.

[21] A. Lozano and N. Jindal, "Transmit diversity vs. spatial multiplexing in modern MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 9, no. 1, pp. 186–197, Jan. 2010, doi: 10.1109/TWC.2010.01.081381.

[22] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010, doi: 10.1109/TWC.2010.092810.091092.

[23] M. Joham, W. Utschick, and J. Nossek, "Linear transmit processing in MIMO communications systems," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 2700–2712, Aug. 2005, doi: 10.1109/TSP.2005.850331.

[24] H. Weingarten, Y. Steinberg, and S. S. Shamai, "The capacity region of the Gaussian multiple-input multiple-output broadcast channel," *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 3936–3964, Sep. 2006, doi: 10.1109/TIT.2006.880064.

[25] R. W. Heath, N. Gonzalez-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 436–453, Apr. 2016, doi: 10.1109/JSTSP.2016.2523924.

[26] R. Méndez-Rial, C. Rusu, N. González-Prelcic, A. Alkhateeb, and R. W. Heath, "Hybrid MIMO architectures for millimeter wave communications: Phase shifters or switches?" *IEEE Access*, vol. 4, pp. 247–267, Jan. 2016, doi: 10.1109/ACCESS.2015.2514261.

[27] C. Mollén, J. Choi, E. G. Larsson, and R. W. Heath, "Achievable uplink rates for massive MIMO with coarse quantization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2017, pp. 6488–6492, doi: 10.1109/ICASSP.2017.7953406.

[28] Y. Li, C. Tao, G. Seco-Granados, A. Mezghani, A. L. Swindlehurst, and L. Liu, "Channel estimation and performance analysis of one-bit massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 65, no. 15, pp. 4075–4089, Aug. 2017, doi: 10.1109/TSP.2017.2706179.

[29] S. Jacobsson, G. Durisi, M. Coldrey, and C. Studer, "Linear precoding with low-resolution DACs for massive MU-MIMO-OFDM downlink," *IEEE Trans. Wireless Commun.*, vol. 18, no. 3, pp. 1595–1609, Mar. 2019, doi: 10.1109/TWC.2019.2894120.

[30] N. Shlezinger and Y. C. Eldar, "Task-based quantization with application to MIMO receivers," *Commun. Inf. Syst.*, vol. 20, no. 2, pp. 131–162, 2020, doi: 10.4310/CIS.2020.v20.n2.a3.

[31] K. Roth, H. Pirzadeh, A. L. Swindlehurst, and J. A. Nossek, "A comparison of hybrid beamforming and digital beamforming with low-resolution ADCs for multiple users and imperfect CSI," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 3, pp. 484–498, Jun. 2018, doi: 10.1109/JSTSP.2018.2813973.

[32] Y. C. Eldar, *Sampling Theory: Beyond Bandlimited Systems*. Cambridge, U.K.: Cambridge Univ. Press, 2015.

[33] Z. Gao, L. Dai, S. Han, I. Chih-Lin, Z. Wang, and L. Hanzo, "Compressive sensing techniques for next-generation wireless communications," *IEEE Wireless Commun.*, vol. 25, no. 3, pp. 144–153, Jun. 2018, doi: 10.1109/MWC.2017.1700147.

[34] L. Liu, E. G. Larsson, W. Yu, P. Popovski, C. Stefanovic, and E. de Carvalho, "Sparse signal processing for grant-free massive connectivity: A future paradigm for random access protocols in the internet of things," *IEEE Signal Process. Mag.*, vol. 35, no. 5, pp. 88–99, Sep. 2018, doi: 10.1109/MSP.2018.2844952.

[35] J. Yuan, Q. He, M. Matthaiou, T. Q. S. Quek, and S. Jin, "Toward massive connectivity for IoT in mixed-ADC distributed massive MIMO," *IEEE Internet Things J.*, vol. 7, no. 3, pp. 1841–1856, Mar. 2020, doi: 10.1109/JIOT.2019.2957281.

[36] V. Monga, Y. Li, and Y. C. Eldar, "Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing," *IEEE Signal Process. Mag.*, vol. 38, no. 2, pp. 18–44, Mar. 2021, doi: 10.1109/MSP.2020.3016905.

[37] N. Samuel, T. Diskin, and A. Wiesel, "Learning to detect," *IEEE Trans. Signal Process.*, vol. 67, no. 10, pp. 2554–2564, May 2019, doi: 10.1109/TSP.2019.2899805.

[38] L. Pellaco, M. Bengtsson, and J. Jaldén, "Matrix-inverse-free deep unfolding of the weighted MMSE beamforming algorithm," *IEEE Open J. Commun. Soc.*, vol. 3, pp. 65–81, 2022, doi: 10.1109/OJCOMS.2021.3139858.

[39] Ö. T. Demir, E. Björnson, and L. Sanguinetti, "Foundations of user-centric cell-free massive MIMO," *Found. Trends® Signal Process.*, vol. 14, nos. 3–4, pp. 162–472, Jan. 2021, doi: 10.1561/2000000109.

[40] J. A. Zhang et al., "An overview of signal processing techniques for joint communication and radar sensing," *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 6, pp. 1295–1315, Nov. 2021, doi: 10.1109/JSTSP.2021.3113120.

[41] E. Björnson, H. Wymeersch, B. Matthiesen, P. Popovski, L. Sanguinetti, and E. de Carvalho, "Reconfigurable intelligent surfaces: A signal processing perspective with wireless applications," *IEEE Signal Process. Mag.*, vol. 39, no. 2, pp. 135–158, Mar. 2022, doi: 10.1109/MSP.2021.3130549.

SP