This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2646659, IEEE Transactions on Signal Processing

On the Quantization of Phase Shifters for Hybrid Precoding Systems

Yuan-Pei Lin
Dept. Electrical Engineering
National Chiao Tung University, Hsinchu, Taiwan
Tel/Fax: 886-3-5731632, Email: ypl@mail.nctu.edu.tw

***Abstract*—Hybrid precoding that combines analog RF processing and digital baseband processing has been shown to be a promising technique for transceivers with limited number of RF chains. The RF processing, due to power and complexity issue, is typically done using phase shifters. A recent study shows that hybrid precoding is no loss of generality if each coefficient of the analog precoder is implemented using two phase shifters. In this paper we consider two possible implementation structures that use Two pHase shIfters for each Coefficient (THIC). These two structures, having the same performance when the phases can take on continuum values, behave differently upon quantization, as the analysis shows. For a small number of quantization bits, the precoder implemented using THIC is a good approximation of the unquantized one. With 3-bit phase shifters, the performance is close to that of the unquantized system in multiuser communications, where the performance is more sensitive to quantization errors. While the analysis in this paper is done assuming high-resolution quantization, simulations show that the result is accurate even for coarse quantization.**

I. Introduction

The performance of a MIMO system is known to improve with the number of antennas on the two transmission ends. Recent advances show that it is feasible to pack a large number of antennas in a small area, particularly in millimeter wave (mmWave) communication systems that use small wavelengths. However cost and power constraints often prohibit having one dedicated RF (radio frequency) chain for each antenna [1][2]. Innovative techniques have been proposed in the literature to overcome the RF limitation.

A promising technique is the so called hybrid precoding scheme, in which analog processing of RF signals is combined with digital processing in the baseband to improve the performance within the RF chain constraint. Analog RF processing, due to power and complexity consideration, is typically implemented using phase shifters [3][4] and the elements of the analog precoder are of unity magnitude. It has been shown that having two RF chains is sufficient to achieve the full beamforming gain [5]. Exploiting the sparsity nature of mmWave channels, the analog precoder proposed in [3] employs column vectors that correspond to some dominating transmission paths, appropriately chosen using orthogonal matching pursuit. The digital precoder can further enhance the system performance when more RF chains are available. Based on the work of [5], it is further shown in [6] (and also independently in [7]) that for the transmission of N_s substreams, the number of RF chains needed for achieving the full multiplexing gain is just N_s. This is done by implementing each coefficient of the optimal unconstrained precoder using two phase shifters and the digital precoder does nothing but scaling each substream properly.

In practical applications, the phase shifters are controlled digitally and assume a finite number of values, depending on the quantization bits. Quantization of the phases can cause a significant degradation to the system performance [24][9]. The high cost of high-resolution phase shifters has motivated the design of systems that use one-bit phase shifters [24][10]. A genetic algorithm is proposed in [11] to design beamformers using low-resolution phase shifters. A heuristic design of the RF precoder with quantized phase shifters is proposed in [12] by designing the precoder iteratively. In [13], vector quantization technique is employed to design codebooks for the RF precoder. Quantization effect of phase shifters on mmWave beamforming antenna arrays is evaluated in [14].

In this paper we consider quantization of phase shifters for the RF precoder in a hybrid precoding scheme. Two possible implementation structures that use Two pHase shIfters for each Coefficient (THIC) of the analog RF precoder are considered. These implementation structures achieve the full multiplexing gain when the phase shifters are not quantized, but they behave differently in the presence of quantization. We analyze the mean square quantization errors (MSQE) of the two structures. With the second structure, the quantization errors of individual phase shifters do not contribute to the overall quantization error in the same manner. The performance benefits from appropriate allocation of quantization bits. As a result one of the two phase shifters can be of a lower resolution and achieve a performance comparable to that of the first structure. For the case the number of substreams transmitted is one, i.e., beamforming case, we derive a lower bound of the SNR degradation due to quantization. The bound, establishing a connection between the quantization error and the resulting SNR degradation, allows us to give an estimate of the SNR degradation based on MSQE. Although the analysis is done under the assumption of high-rate quantization, simulations show that the results are accurate even in coarse quantization. Examples show that with an average of two-bit quantization, the quantized THIC hybrid precoder that employs the minimum number of RF chains is less than 1 dB away from the optimal unconstrained system that is endowed with one RF chain

This work was supported by Ministry of Science and Technology, Taiwan, R. O. C., under MOST 104-2221-E-009 -080 -MY2.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2646659, IEEE Transactions on Signal Processing

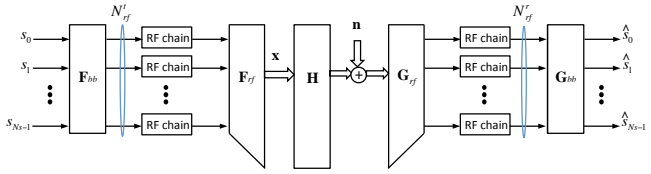


Fig. 1. A MIMO communication system with hybrid precoding.

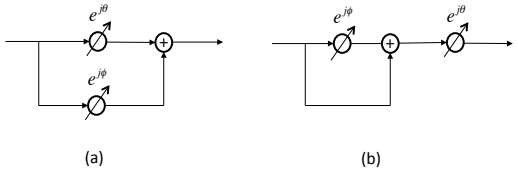


Fig. 2. Phase shifter implementation, (a) Structure I, and (b) Structure II.

per antenna. In comparison to a conventional RF precoder that uses one phase shifter per coefficient, only half as many RF chains are needed. Furthermore, with 3-bit phase shifters, the performance is close to that of the unquantized system in multiuser communications, where the performance is more sensitive to quantization errors.

Notation. The variance of a random variable x is denoted as σ_x^2. The 2-norm of a vector $\mathbf{f}$ is denoted as $||\mathbf{f}||$ and the k-th entry of $\mathbf{f}$ is denoted as $[\mathbf{f}]_k$. The notation $\mathbf{A}^\dagger$ denotes the transpose and conjugate of a matrix $\mathbf{A}$. The expectation of a random variable x is denoted by $E[x]$.

II. System model

Consider the wireless system with N_t transmit antennas and N_r receive antennas in Fig. 1. The channel is modelled by an $N_r \times N_t$ matrix $\mathbf{H}$ with $N_r \times 1$ channel noise $\mathbf{n}$. We assume the channel is slow fading so that the channel does not change during each channel use. The noise vector $\mathbf{n}$ is assumed to be additive white Gaussian with zero mean and variance σ_n^2. Suppose the numbers of RF chains at the transmitter and the receiver are, respectively, N^t_{rf} and N^t_{rf} and the number of substreams transmitted is N_s. The input of the transmitter is a $N_s \times 1$ vector $\mathbf{s}$, whose elements $s_0, s_1, \cdots, s_{N_s-1}$ are modulations symbols that are assumed to be uncorrelated, of zero mean and of variance $\sigma^2_{s_0}, \sigma^2_{s_1}, \cdots, \sigma^2_{s_{N_s-1}}$.

The transmit precoding matrix $\mathbf{F}$, of dimensions $N_t \times N_s$, is the product $\mathbf{F} = \mathbf{F}_{rf}\mathbf{F}_{bb}$. The matrix $\mathbf{F}_{bb}$, of size $N^t_{rf} \times N_s$, represents the baseband processing while $\mathbf{F}_{rf}$, of size $N_t \times N^t_{rf}$, corresponds to the RF signal processing that consists of phase shifters. Although $\mathbf{F}_{rf}$ can be implemented using phase shifters, its entries are not necessarily of constant magnitude. For example, the entries can be the sums of two phase shifters. Similarly, the receive matrix is of the form $\mathbf{G} = \mathbf{G}_{bb}\mathbf{G}_{rf}$, where $\mathbf{G}_{bb}$ is $N_s \times N^r_{rf}$ and it represents the baseband processing at the receiver. The matrix $\mathbf{G}_{rf}$, of size $N^r_{rf} \times N_r$, corresponds to the RF signal processing. Like $\mathbf{F}_{rf}$, the matrix $\mathbf{G}_{rf}$ is implemented using phase shifters, but its entries are not necessarily of constant magnitude.

The transmitter output is $\mathbf{x} = \mathbf{F}_{rf}\mathbf{F}_{bb}\mathbf{s}$. As the input modulation symbols of the transmitter are uncorrelated and of zero mean, the total transmit power is $P_t = \sum_{i=0}^{N_s-1} ||\mathbf{f}_i||^2\sigma^2_{s_i}$, where $\mathbf{f}_i$ is the i-th column vector of the precoder $\mathbf{F}$. The receiver output vector is $\widehat{\mathbf{s}} = \mathbf{Ts}+\mathbf{e}$, where $\mathbf{T} = \mathbf{GHF}$ and $\mathbf{e} = \mathbf{Gn}$. When the receiver is zero-forcing and $\mathbf{T} = \mathbf{I}_{N_s}$, the number of bits that can be transmitted is given by

$$\mathcal{R}_b = \sum_{i=0}^{N_s-1} \log_2\left(1 + \frac{\sigma^2_{s_i}}{\Gamma\sigma^2_{e_i}}\right), \tag{1}$$

where $\sigma^2_{e_i}$ is the variance of the i-th entry of the noise term $\mathbf{e}$ and Γ is the SNR gap that depends on the desired error rate, e.g., $\Gamma = -\ln(5BER)/1.5$ [15].

III. Phase shifter implementation of precoder coefficients

In most earlier designs, the analog RF precoder is constrained to have constant magnitude due to phase shifter implementation. It is shown recently that the coefficients of the analog RF precoder can be of non-constant magnitude by using two phase shifters for each coefficient [6]. In particular, it is observed that a complex number $c = re^{j\alpha}$ with $0 \le r \le 2$ can be expressed as $c = e^{j\theta} + e^{j\phi}$, where

$$\theta = \alpha + \cos^{-1}(r/2), \quad \phi = \alpha - \cos^{-1}(r/2).$$

The result suggests that the multiplication of a number by a scalar c with magnitude ≤ 2 can be implemented as Structure I in Fig. 2(a).[1] When the phases θ and ϕ are quantized to a finite set of values, so is c. Alternatively, we can express c as

$$c = e^{j\theta}(1 + e^{j\phi}), \tag{2}$$

where θ and ϕ are now related to the magnitude and phase of c in a different manner. The expression in (2) gives rise to Structure II in Fig. 2(b), which also uses two phase shifters and a combiner. The two structures in Fig. 2 yield the same result when θ and ϕ are not quantized. In the presence of quantization they have different performance as to be shown in the following analysis.

Let us consider the quantization of a complex scalar $c = re^{j\alpha}$ with $r \le 2$ using the two structures in Fig. 2 by quantizing the phases θ and ϕ. The following assumptions are made in the quantization of θ and ϕ.

- A1. The quantization errors of θ and ϕ are uncorrelated. That is $\delta_\theta = \hat{\theta} - \theta$ and $\delta_\phi = \hat{\phi} - \phi$ are uncorrelated, where $\hat{\theta}$ and $\hat{\phi}$ are respectively the quantized values of θ and ϕ.
- A2. The quantization error δ_θ is assumed to be uncorrelated with θ, of zero mean, and uniformly distributed over the interval $[-\Delta_\theta/2, \Delta_\theta/2)$, where Δ_θ is the quantization step size. The same assumption is made on the quantization error δ_ϕ.

These assumptions, commonly used in the analysis of quantization error [16], is generally valid for high-rate quantization. Let θ be drawn from an interval of length V_θ

[1] In actual implementation the power at the output is not higher than that at the input when only passive devices are used. To reflect this fact, it is more appropriate to do a scaling of $1/2$ for each branch after the splitter. The scaling does not affect the subsequent analysis on coefficient quantization and it is not shown in the figure.

and the number of bits for quantizing θ be b_θ, then the quantization step size is given by $\Delta_\theta = V_\theta 2^{-b}$. In this case the variance of quantization error $\sigma^2_{\delta_\theta} = E[\delta^2_\theta]$ can be given in a closed form by

$$\sigma^2_{\delta_\theta} = \frac{1}{12}\Delta^2_\theta = \frac{V^2_\theta}{12}2^{-2b_\theta}. \tag{3}$$

Although high-rate quantization is assumed in the derivations, we will see in the simulation examples that the result is accurate even when the quantization resolution is low.

A. Quantization: Structure I

In Structure I, a scalar $c = re^{j\alpha}$ is expressed as $e^{j\theta}+e^{j\phi}$. When θ and ϕ are quantized to $\hat\theta$ and $\hat\phi$, correspondingly c is quantized to $\hat c = e^{j\hat\theta}+e^{j\hat\phi}$. The quantization error $\hat c - c$ is not related to the quantization of θ and ϕ in a straight forward manner. However when the quantization resolution is high, the mean square quantization error (MSQE) $\mathcal{E}_1 = E[|\hat c - c|^2]$ can be approximated in terms of those of θ and ϕ.

Lemma 1. *Consider a random variable $c = e^{j\theta} + e^{j\phi}$ that is quantized by quantizing θ and ϕ. With high-resolution quantization, the effective MSQE for Structure I can be approximated as*

$$\mathcal{E}_1 \approx \sigma^2_{\delta_\theta} + \sigma^2_{\delta_\phi}, \tag{4}$$

where $\sigma^2_{\delta_\theta}$ and $\sigma^2_{\delta_\phi}$ are, respectively, the variances of the quantization errors δ_θ and δ_ϕ.

A proof of Lemma 1 can be found in Appendix A. For $0 \le r \le 2$ and $0 \le \alpha < 2\pi$, the ranges of θ and ϕ are both from 0 to 2π. Thus for both cases the intervals of quantization are of lengths 2π and the step sizes for quantizing θ and ϕ are $\Delta_\theta = 2\pi 2^{-b_\theta}$ and $\Delta_\phi = 2\pi 2^{-b_\phi}$, where b_θ and b_ϕ are, respectively, the number of bits used for quantizing θ and ϕ. We obtain the effective error expression $\mathcal{E}_1 = \frac{\pi^2}{3}(2^{-2b_\theta} + 2^{-2b_\phi})$. In this case, we see the quantization errors of θ and ϕ contribute equally to the overall error. Let the average number of quantization bits be $b = (b_\theta + b_\phi)/2$. For a given b, we choose $b_\theta = b_\phi = b$ and we arrive at $\mathcal{E}_1 = \frac{2\pi^2}{3}2^{-2b}$.

B. Quantization: Structure II

An arbitrary scalar $c = re^{j\alpha}$ with $r \le 2$ can also be written as $c = e^{j\theta}(1 + e^{j\phi})$, where

$$\theta = \alpha - \cos^{-1}(r/2), \quad \phi = 2\cos^{-1}(r/2). \tag{5}$$

Thus we have the implementation as in Structure II (Fig. 2(b)). Conversely we can express the magnitude and phase of c as

$$r = 2\cos(\phi/2), \quad \alpha = \theta + \phi/2. \tag{6}$$

When θ and ϕ are quantized to $\hat\theta$ and $\hat\phi$, correspondingly c is quantized to $\hat c = e^{j\hat\theta}(1+e^{j\hat\phi})$. The quantization error is given by $\hat c - c = e^{j\hat\theta}(1+e^{j\hat\phi}) - e^{j\theta}(1+e^{j\phi})$, which is not directly related to the quantization errors of θ and ϕ. Assume r, δ_θ, and δ_ϕ are uncorrelated, a reasonable assumption when the quantization resolution is high. Like Structure I, the MSQE $\mathcal{E}_2 = E[|\hat c - c|^2]$ can be approximated in terms of those of θ and ϕ.

Lemma 2. *Consider the quantization of a random variable $c = re^{j\alpha}$ with $r \le 2$. Express c as $c = e^{j\theta}(1 + e^{j\phi})$ and quantize c by quantizing θ and ϕ. When r, δ_θ, and δ_ϕ are uncorrelated, and the quantization resolution is high, the effective MSQE $\mathcal{E}_2$ for Structure II can be approximated as*

$$\mathcal{E}_2 \approx E[r^2]\sigma^2_{\delta_\theta} + \sigma^2_{\delta_\phi}. \tag{7}$$

A proof is give in Appendix B. For $0 \le r \le 2$ and $0 \le \alpha < 2\pi$, the range of θ is from 0 to 2π. But unlike Structure I, the range of ϕ now is from 0 to only π. This is because $\phi = 2\cos^{-1}(r/2)$ from (6) and r is a nonnegative number. Thus the lengths of quantization $V_\theta = 2\pi$ and $V_\phi = \pi$, and the step sizes for quantizing θ and ϕ are, respectively, $\Delta_\theta = 2\pi 2^{-b_\theta}$ and $\Delta_\phi = \pi 2^{-b_\phi}$. As a result, the variances of δ_θ and δ_ϕ are, respectively,

$$\sigma^2_{\delta_\theta} = \frac{\pi^2}{3}2^{-2b_\theta}, \quad \sigma^2_{\delta_\phi} = \frac{\pi^2}{12}2^{-2b_\phi}. \tag{8}$$

We obtain the effective quantization error

$$\mathcal{E}_2 = \frac{\pi^2}{3}\left(E[r^2]2^{-2b_\theta} + \frac{1}{4}2^{-2b_\phi}\right). \tag{9}$$

We see that the quantization errors of θ and ϕ do not contribute to the overall error in the same way. Suppose the average number of bits that can be used for quantizing θ and ϕ is b, and we are allowed to allocate the bits between θ and ϕ. Suppose b_θ bits are used for quantizing θ and b_ϕ bits for ϕ. Then the bit allocation problem becomes

$$\min_{b_\theta, b_\phi, \text{subject to } (b_\theta+b_\phi)/2=b} \mathcal{E}_2 = \frac{\pi^2}{3}\left(E[r^2]2^{-2b_\theta} + \frac{1}{4}2^{-2b_\phi}\right).$$

The bit allocation problem can be solved using arithmetic mean–geometric mean inequality [17], in particular $\mathcal{E}_2 \ge \frac{\pi^2}{3}2^{-2b}\sqrt{E[r^2]}$, where we have used the constraint $b_\theta + b_\phi = 2b$. The lower bound is independent of b_θ and b_ϕ; the optimal bit allocation is such that the inequality becomes an equality, i.e., the two terms $E[r^2]2^{-2b_\theta}$ and $\frac{1}{4}2^{-2b_\phi}$ are of the same value. Thus the optimal bit allocation is

$$b_\theta = b + \frac{1}{2}\log_2\left(2\sqrt{E[r^2]}\right), \quad b_\phi = b - b_\theta,$$

and the resulting MSQE is $\mathcal{E}_2 = \frac{\pi^2}{3}2^{-2b}\sqrt{E[r^2]}$. The above equation in general yields non integer solution, which can be rounded to integers in practical implementation.

Joint scalar quantization. In the above derivation of $\mathcal{E}_2$, the two phases θ and ϕ are quantized separately. One is quantized without consideration of the other. Observe that, instead of (5) we can also determine θ and ϕ using $\phi = 2\cos^{-1}(r/2)$ and $\theta = \alpha - \phi/2$. If we first quantize ϕ to $\hat\phi$, the desired unquantized value of θ is now changed to $\alpha - \hat\phi/2$, which can then be quantized to obtain $\hat\theta$. Such an approach, still using scalar quantization, quantizes θ and ϕ jointly and has a smaller quantization error as we see next. Let the quantized θ be $\hat\theta = \alpha - \hat\phi/2 + \delta_\theta$. Thus $\hat c = e^{j(\hat\theta+\hat\phi/2)}2\cos(\hat\phi/2)$ can be expressed as $\hat c = e^{j(\alpha+\delta_\theta)}2\cos(\hat\phi/2)$. Now the equivalent phase is affected only by the quantization error δ_θ rather than both δ_θ and δ_ϕ as in the previous case. When θ and ϕ are thus quantized, the error can be approximated as (a proof given in Appendix C.)

$$\mathcal{E}'_2 \approx E[r^2]\sigma^2_{\delta_\theta} + (1 - E[r^2]/4)\sigma^2_{\delta_\phi}, \tag{10}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2646659, IEEE Transactions on Signal Processing

assuming high-resolution quantization. The MSQE in (10) is smaller than that in (9) by $E[r^2]/4\sigma_{\delta_\phi}^2$. The ranges of θ and ϕ are the same as before. Combining (8) and (10), we obtain

$$\mathcal{E}_2' = \frac{\pi^2}{3}\left(E[r^2]2^{-2b_\theta} + \frac{1}{16}(4 - E[r^2])2^{-2b_\phi}\right). \quad (11)$$

Note that in such a joint quantization of θ and ϕ, only scalar quantization is needed. Quantization is done the same way as discussed earlier except for the computation of the desired θ.

Uniform quantization of magnitude. With Structure II, the quantized magnitude $\hat{r}$ is determined by the quantized phase $\hat{\phi}$. When a uniform quantizer is used for quantizing ϕ, the effective quantizer for the magnitude is not uniform as $\hat{r} = 2\cos(\hat{\phi}/2)$; there are more reconstruction points around 2 than around 0. If we can choose the reconstruction points for ϕ non-uniformly[2], the effective quantizer for r can be made uniform. In particular, we divide the interval $[0, 2]$ into 2^{b_ϕ} quantization bins, each of length $2/2^{b_\phi}$. Say the center of the bins are $r_1, r_2, \cdots, r_{2^{b_\phi}}$. We choose the reconstruction points for ϕ to be $\{2\cos^{-1}(r_i/2)\}_{i=1}^{2^{b_\phi}}$. With uniform quantization of the magnitude, we can obtain the mean square quantization error of r as $\sigma_{\delta_r}^2 = (1/3)2^{-2b_\phi}$ using the formula in (3) for high-resolution quantization. When the joint scalar quantization discussed above is employed, the error now becomes (a proof given in Appendix D)

$$\mathcal{E}_2'' = \frac{\pi^2}{3}E[r^2]2^{-2b_\theta} + \frac{1}{3}2^{-2b_\phi}. \quad (12)$$

Note that we can not do the same thing with Structure I. For Structure I we have $\hat{r} = 2\cos(\hat{\theta}/2 - \hat{\phi}/2)$, which depends on both $\hat{\theta}$ and $\hat{\phi}$ rather than a single phase. Therefore we do not have direct control over the magnitude of the reconstruction points like Structure II.

IV. Quantization: beamforming systems

In this section, we consider the SNR loss of a beamforming system due to the quantization of the beamformers based on the MSQE derived in the previous section. When there is only one RF chain, $N_{rf}^t = N_{rf}^r = 1$, only one substream can be transmitted and $N_s = 1$. The system in Fig. 1 becomes a beamforming system. In this case, the precoder $\mathbf{F} = \mathbf{F}_{rf} = \mathbf{f}$ is an $N_t \times 1$ beamforming vector and $\mathbf{G} = \mathbf{G}_{rf} = \mathbf{g}$ is a $1 \times N_r$ combining vector. The transmission power is $P_t = \sigma_s^2||\mathbf{f}||^2$. Let the singular value decomposition of the channel be

$$\mathbf{H} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\dagger, \quad (13)$$

where $\mathbf{U}$ and $\mathbf{V}$ are unitary, of sizes $N_r \times N_r$ and $N_t \times N_t$, respectively. The diagonal elements of $\mathbf{\Lambda}$ are in non increasing order, i.e., $\lambda_0 \geq \lambda_1 \geq \cdots \geq \lambda_{K-1}$, where $K = \min\{N_t, N_r\}$. Then the optimal beamforming vector is $\mathbf{f} = \mathbf{v}_0$, where $\mathbf{v}_0$ is the first column vector of $\mathbf{V}$, and the optimal combining vector $\mathbf{g} = \mathbf{u}_0^\dagger$, where $\mathbf{u}_0$ is the first column vector of $\mathbf{U}$. The optimal overall signal to noise ratio is

$$SNR = \lambda_0^2 P_t/\sigma_n^2. \quad (14)$$

[2]To implement a phase shifter with nonuniform phases, one can use the combination of constant-phase phase shifters and switches as detailed in [18].

A. Quantization of beamforming vectors

When there is only one RF chain, baseband processing plays no part and beamforming is done entirely using RF phase shifters. When the phase shifters are of finite resolution, the transmit beamformer and receive combiner are only quantized versions of the optimal beamformer and combiner. Given the quantized beamformer $\hat{\mathbf{f}}$, the optimal unconstrained combining vector is no longer the left singular vector of $\mathbf{H}$ corresponding to the largest singular value. It is the one that is matched to $\mathbf{H}\hat{\mathbf{f}}$. Let $\hat{\mathbf{g}}$ be the quantized combiner. The receiver output is $\hat{s} = \hat{\mathbf{g}}\mathbf{H}\hat{\mathbf{f}}s + \hat{\mathbf{g}}\mathbf{n}$. The overall SNR in this case is given by

$$SNR_{t,r} = \frac{|\hat{\mathbf{g}}\mathbf{H}\hat{\mathbf{f}}|^2}{||\hat{\mathbf{f}}||^2||\hat{\mathbf{g}}||^2}\frac{P_t}{\sigma_n^2}, \quad (15)$$

where the subscript t, r is added to indicate that both transmit and receive sides are quantized. In what follows, we summarize the steps for designing the quantized beamformer and combiner. (1) For a given channel $\mathbf{H}$, we compute the right singular vector $\mathbf{v}_0$ that corresponds to the largest singular value. (2) Normalize the optimal beamformer as $\mathbf{f} = \beta_f\mathbf{v}_0$, where β_f is a positive scalar such that the entries of $\mathbf{f}$ have magnitude at almost 2. (3) Quantize each entry of $\mathbf{f}$ using one of the two structures discussed in Sec. III. Let the kth entry of $\mathbf{f}$ be $f_k = r_k e^{j\alpha_k}$. With Structure I, f_k is quantized to $\hat{f}_k = e^{j\hat{\theta}_k} + e^{j\hat{\phi}_k}$, where $\hat{\theta}_k$ and $\hat{\phi}_k$ are respectively the quantized values of

$$\theta_k = \alpha_k + \cos^{-1}(r_k/2), \quad \phi_k = \alpha_k - \cos^{-1}(r_k/2). \quad (16)$$

With Structure II, f_k is quantized to $[\hat{\mathbf{f}}]_k = e^{j\hat{\theta}_k}(1 + e^{j\hat{\phi}_k})$, where $\hat{\theta}_k$ and $\hat{\phi}_k$ are respectively the quantized values of

$$\phi_k = 2\cos^{-1}(r_k/2), \quad \theta_k = \alpha_k - \phi_k/2, \quad (17)$$

for joint scalar quantization of θ_k and ϕ_k. (4) Compute the desired combiner $(\mathbf{H}\hat{\mathbf{f}})^\dagger$ and apply normalization $\mathbf{g} = \beta_g(\mathbf{H}\hat{\mathbf{f}})^\dagger$, where $\beta_g \geq 0$ is such that the entries of $\mathbf{g}$ have magnitude ≤ 2. (5) Quantize $\mathbf{g}$ as in Step (3).

For each coefficient of the beamforming vector, two phase shifters are needed when either Structure I or II is used. A total of $2N_t$ phase shifters are needed for implementing the transmit beamformer and $2N_r$ phase shifters needed for implementing the receive combiner.

B. SNR degradation due to quantization

With the quantization of the transmit beamformer and the receive combiner, the SNR is degraded. There is more degradation when the vectors are more heavily quantized. It turns out that a connection between the quantization error and the resulting SNR degradation can be established. An estimate of SNR loss can obtained based on the quantization error, irrespective of the quantization scheme adopted.

We first consider the case when only the transmit beamforming vector $\mathbf{f}$ is quantized but not the combiner. Choosing the unconstrained combiner as $(\mathbf{H}\hat{\mathbf{f}})^\dagger$, then the SNR at the combiner output is $SNR_t = \frac{||\mathbf{H}\hat{\mathbf{f}}||^2}{||\hat{\mathbf{f}}||^2}\frac{P_t}{\sigma_n^2}$, where a subscript t is used to indicate that the transmit beamformer is quantized. In comparison with the SNR without quantization

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2646659, IEEE Transactions on Signal Processing

in (14), the SNR degradation $\mathcal{D}_t = SNR/SNR_t$ is given by

$$\mathcal{D}_t = \frac{\lambda_0^2||\hat{\mathbf{f}}||^2}{||\mathbf{H}\hat{\mathbf{f}}||^2}.$$

When both the beamformer and combiner are quantized, the SNR is as given in (15). In this case, the SNR degradation $\mathcal{D}_{t,r} = SNR/SNR_{t,r}$ is

$$\mathcal{D}_{t,r} = \frac{\lambda_0^2||\hat{\mathbf{f}}||^2||\hat{\mathbf{g}}||^2}{|\hat{\mathbf{g}}\mathbf{H}\hat{\mathbf{f}}|^2}. \quad (18)$$

To analyze the effect of quantization on the SNR degradation, suppose the quantized beamformer is $\hat{\mathbf{f}} = \mathbf{f} + \mathbf{d}_f$, where $\mathbf{d}_f$ is the quantization error vector, whose entires are assumed to be uncorrelated random variables with zero mean and variance $\sigma_{d_f}^2$. When the receiver side is also quantized, suppose the desired normalized combiner is quantized to $\hat{\mathbf{g}} = \mathbf{g}+\mathbf{d}_g$, where $\mathbf{d}_g$ is the quantization error vector whose entires are assumed to be uncorrelated random variables with zero mean and variance $\sigma_{d_g}^2$.

Lemma 3. *Consider the case the quantization resolution is high and the transmit beamformer is quantized but not the receive combiner. As the number of transmit antennas N_t tends to infinity, the average SNR degradation satisfies*

$$\mathcal{D}_t \approx 1 + ||\mathbf{d}_f||^2]/||\mathbf{f}||^2 - ||\mathbf{H}\mathbf{d}_f||^2/(\lambda_0^2||\mathbf{f}||^2). \quad (19)$$

For the case both beamformer and combiner are quantized, we have

$$\mathcal{D}_{t,r} \approx \mathcal{D}_t(1 + ||\mathbf{d}_g||^2/||\mathbf{g}||^2), \quad (20)$$

as N_t and N_r tend to infinity.

A proof can be found in Appendix E. The lemma makes a connection between the quantization error and the SNR degradation of the beamforming system. Such a connection can be further used to quantify the average SNR degradation. Notice that the last term on the right hand side of (19) is equal to $\sum_{k=0}^{N_t-1} \lambda_i^2/\lambda_0^2|[\mathbf{V}\mathbf{d}_f]_k|^2/||\mathbf{f}||^2$. Each term in the summation contains the factor λ_i^2/λ_0^2. Typically the first few eigenvalues are significantly larger than the rest for a large N_t and so $\sum_{k=0}^{N_t-1} \lambda_i^2/\lambda_0^2|[\mathbf{V}\mathbf{d}_f]_k|^2/||\mathbf{f}||^2 \ll \sum_{k=0}^{N_t-1} |[\mathbf{V}\mathbf{d}_f]_k|^2/||\mathbf{f}||^2$, which is equal to $||\mathbf{d}_f||^2/||\mathbf{f}||^2$ as $\mathbf{V}$ is unitary. Therefore the last term on the right hand side of (19) is much smaller than the second term. We have the approximation $\mathcal{D}_t \approx 1 + ||\mathbf{d}_f||^2]/E[||\mathbf{f}||^2$ and thus $E[\mathcal{D}_t] \approx 1 + E[||\mathbf{d}_f||^2/||\mathbf{f}||^2]$. For high-resolution quantization, it is reasonable to assume that $||\mathbf{d}_f||^2$ and $1/||\mathbf{f}||^2$ are correlated. This leads to

$$E[\mathcal{D}_t] \gtrsim \bar{\mathcal{D}}_t, \text{ where } \bar{\mathcal{D}}_t = 1 + E[||\mathbf{d}_f||^2]/E[||\mathbf{f}||^2], \quad (21)$$

where the inequality is due to Jensen's inequality for a convex function and $E[1/x] \geq 1/E[x]$. For the case that both the beamformer and the combiner are quantized, apply the approximation $\mathcal{D}_t \approx 1 + ||\mathbf{d}_f||^2/||\mathbf{f}||^2]$ and (20) becomes $\mathcal{D}_{t,r} \approx 1 + ||\mathbf{d}_f||^2/||\mathbf{f}||^2 + ||\mathbf{d}_g||^2/||\mathbf{g}||^2 + ||\mathbf{d}_f||^2||\mathbf{d}_g||^2/(||\mathbf{f}||^2||\mathbf{g}||^2)$. Ignoring the last term, which is a lot smaller then the first three, we get $\mathcal{D}_{t,r} \gtrsim 1 + ||\mathbf{d}_f||^2/||\mathbf{f}||^2+||\mathbf{d}_g||^2/||\mathbf{g}||^2$. Taking the expectation of $\mathcal{D}_{t,r}$ and using Jensen's inequality as in the above derivation for $E[\mathcal{D}_t]$ we arrive at

$$E[\mathcal{D}_{t,r}] \gtrsim \bar{\mathcal{D}}_{t,r}, \quad (22)$$

where $\bar{\mathcal{D}}_{t,r} = 1+E[||\mathbf{d}_f||^2]/E[||\mathbf{f}||^2]+E[||\mathbf{d}_g||^2]/E[||\mathbf{g}||^2]$.

The bounds $\bar{\mathcal{D}}_t$ and $\bar{\mathcal{D}}_{t,r}$ can be used to estimate the SNR loss due to quantization. The error term $E[||\mathbf{d}_f||^2]$ is equal to $N_t\sigma_{d_f}^2$ and $E[||\mathbf{d}_g||^2] = N_r\sigma_{d_g}^2$, where $\sigma_{d_f}^2$ and $\sigma_{d_g}^2$ are quantization errors of the coefficients and they can be computed using the results derived in Sec. III. Observe that $\bar{\mathcal{D}}_t$ and $\bar{\mathcal{D}}_{t,r}$ are directly related to the signal to quantization error ratio. Let $SQR_{q,t} = E[||\mathbf{f}||^2]/E[||\mathbf{d}_f||^2]$, the signal to quantization error ratio for the transmit beamformer. Similarly, we can define $SQR_{q,r} = E[||\mathbf{g}||^2]/E[||\mathbf{d}_g||^2]$, the signal to quantization error ratio for the receive combiner. Then $\bar{\mathcal{D}}_t = 1+1/SQR_{q,t}$ and $\bar{\mathcal{D}}_{t,r} = 1+1/SQR_{q,t}+1/SQR_{q,r}$. The computation of $SQR_{q,t}$ requires $E[||\mathbf{f}||^2]$, which can be estimated using simulations or computed if statistics of $||\mathbf{f}||$ are available. For example, assume the modulus of each element of $\mathbf{f}$ is uniformly distributed over $[0\ 2]$, then $E[||\mathbf{f}||^2] = \frac{4}{3}N_t$. Using the MSQE derived for S2 in (12) and applying optimal bit allocation between θ and ϕ, we have $\mathcal{D}_t = 1+2.85\times 2^{-2b}$, which is equal to 1.18 for $b = 2$. For the case when only the receiver beamformer is quantized, the loss can be obtained the same way. That is, we can obtain around 85% of the SNR achieved by the unquantized beamformer when only the transmit (or receive) beamformer is quantized. The can be compared with the results in [19], in which each receiver beamforming coefficient is implemented using one phase shifter and $N_t = 1$, i.e., no quantization of transmit beamformer. It is shown therein that around 66% of the SNR achieved by the unquantized beamformer can be obtained. Therefore when $N_t = 1$, S2 quantization that uses two phase shifters for each beamforming coefficient has an improvement of around 19%. Although a large antenna size is assumed in the derivations, simulations show that the accuracy of the estimate is not greatly affected by the antenna size.

V. Quantization of RF precoders

For the beamforming case, it suffices to use only one RF chain. For the transmission of N_s substreams, we only need as many RF chains for both the transmitting and receiving ends. First we consider the case that the number of substreams is the same as the number of RF chains, i.e., $N_s = N_{rf}^t = N_{rf}^r$. When there is no quantization on the phases shifters, the optimal precoder is $\mathbf{V}_{N_s}$, where $\mathbf{V}_{N_s}$ consists of the first N_s column vectors of $\mathbf{V}$, and $\mathbf{V}$ is the unitary matrix given in (13). The optimal receive matrix is $\mathbf{U}_{N_s}^\dagger$, where $\mathbf{U}_{N_s}$ consists of the first N_s column vectors of $\mathbf{U}$.

The quantization of the precoder is similar to the beamforming case. We normalize and quantize each column vector of the optimal precoder as in Sec. IV-A. In particular, the desired RF precoder is $\mathbf{F}_{rf} = \mathbf{V}_{N_s}\boldsymbol{\beta}_f$, where $\boldsymbol{\beta}_f$ is a diagonal matrix with positive diagonal elements $\beta_{f,1}$, $\beta_{f,2}, \cdots \beta_{f,N_s}$ and $\beta_{f,k}$ is such that the k-th column of $\mathbf{F}_{rf}$ has maximum magnitude equal to 2. Quantize each entry of $\mathbf{F}_{rf}$ using Structures I or II and call the quantized matrix $\widehat{\mathbf{F}}_{rf}$, the analog precoder that can be implemented using quantized phase shifters. The quantized $\widehat{\mathbf{F}}_{rf}$ is different from the optimal precoder, depending on the quantization resolution of the phase shifters. For the residual difference, we can use digital processing $\mathbf{F}_{bb}$ to help with further precoding. When the receiver is not constrained, it is shown in [20]

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2646659, IEEE Transactions on Signal Processing

that, for a given $\mathbf{F}_{rf}$ the optimal $\mathbf{F}_{bb}$ that maximimizes the capacity of the system is given by $(\widehat{\mathbf{F}}_{rf}^{\dagger}\widehat{\mathbf{F}}_{rf})^{-\frac{1}{2}}\mathbf{V}_{eff}$, where $\mathbf{V}_{eff}$ is the $N_{rf}^t \times N_s$ unitary matrix that consists of the right singular vectors of $\mathbf{H}\widehat{\mathbf{F}}_{rf}$. At the receiver side, the zero-forcing receive matrix that minimizes the total noise is the Moore-Penrose left inverse of $\mathbf{HF}$, where $\mathbf{F} = \widehat{\mathbf{F}}_{rf}\mathbf{F}_{bb}$. The row vectors of the left inverse are normalized to obtain the desired $\mathbf{G}_{rf}$. That is $\mathbf{G}_{rf} = \boldsymbol{\beta}_g((\mathbf{HF})^{\dagger}(\mathbf{HF}))^{-1}(\mathbf{HF})^{\dagger}$, where $\boldsymbol{\beta}_g$ is a diagonal matrix with positive diagonal elements $\beta_{g,1}, \beta_{g,2}, \cdots \beta_{g,N_s}$ and $\beta_{g,k}$ is such that the k-th row of $\mathbf{G}_{rf}$ has maximum magnitude equal to 2. Quantize the entries of $\mathbf{G}_{rf}$ using Structure I or II to obtained $\widehat{\mathbf{G}}_{rf}$, which can be implemented using quantized phase shifters. With the quantized $\widehat{\mathbf{G}}_{rf}$, the product $\widehat{\mathbf{G}}_{rf}\mathbf{HF}$ is not the identity matrix. There is some interference among the substreams. To remove the interference, we can choose $\mathbf{G}_{bb}$ to be the inverse of $\mathbf{G}_{rf}\mathbf{HF}$.

In the above discussion, the number of RF chains is the same as the number of substreams. When there are more RF chains than substreams, $\mathbf{F}_{rf}$ has more than N_s columns and there are extra phase shifters. These phase shifters can be used to reduce the quantization error of some of the substreams. For example, suppose the extra phase shifters are used for the first substream, then there are more phase shifters to represent the coefficients of the first column of the $\mathbf{F}_{rf}$ and a higher resolution can be achieved as discussed in Sec. III. In this case the solution of zero-forcing $\mathbf{G}_{bb}$ is not unique. We can choose $\mathbf{G}_{bb}$ to be the left inverse of $\widehat{\mathbf{G}}_{rf}\mathbf{HF}$ that has the smallest total noise power, i.e., the Moore-Penrose inverse $\mathbf{G}_{bb} = ((\widehat{\mathbf{G}}_{rf}\mathbf{HF})^{\dagger}(\widehat{\mathbf{G}}_{rf}\mathbf{HF}))^{-1}(\widehat{\mathbf{G}}_{rf}\mathbf{HF})^{\dagger}$.

VI. Simulations

For the evaluation of the proposed quantization schemes, we adopt the clustered channel representation that is useful for modelling multipath propagation. The channel consists of N_{cl} clusters and each cluster contains N_{ray} propagation paths [21][22],

$$\mathbf{H} = \frac{1}{\sqrt{N_{cl}N_{ray}}} \sum_{k=1}^{N_{cl}} \sum_{\ell=1}^{N_{ray}} \alpha_{k\ell}\mathbf{a}_r(\phi_{a,kl}^r, \phi_{e,kl}^r)\mathbf{a}_t^{\dagger}(\phi_{a,kl}^t, \phi_{e,kl}^t), \tag{23}$$

where $\alpha_{k\ell}$, the complex gain of the ℓth ray in the kth cluster, is assumed to a Gaussian random variable of zero mean and unity variance. The azimuth (elevation) angle of departure $\phi_{a,kl}^t$ ($\phi_{e,kl}^t$) and azimuth (elevation) angle of arrival $\phi_{a,kl}^r$ ($\phi_{e,kl}^r$) are of a truncated Laplacian distribution [23][21]. The means of angles of departure (arrival) in azimuth and elevation are assumed to be uniformly distributed over $[0\ 2\pi]$. The vectors $\mathbf{a}_t^{\dagger}(\phi_{a,kl}^t, \phi_{e,kl}^t)$ and $\mathbf{a}_r(\phi_{a,kl}^r, \phi_{e,kl}^r)$ are, respectively, the transmit and receive antenna array response vectors. The array response for a uniform planar array with $M \times N$ antennas is given by $[\mathbf{a}(\phi_a, \phi_e)]_{m+nM} = e^{j2\pi d(m\cos(\phi_e)+n\sin(\phi_e)\cos(\phi_a))}$, for $0 \le m < M$ and $0 \le n < N$, where d is the antenna spacing normalized by the wavelength. In the simulation examples, we assume the antenna spacing to be half wavelength and $d = 1/2$. The standard deviations of angles of departure (arrival) in azimuth and elevation, also called angular spreads, are assumed to be 7.5°, unless mentioned otherwise, and we use $N_{cl} = 3$ and $N_{ray} = 10$ [21]. The clustered channel model is used Examples 2–5 and 10^5 channels are used in the performance evaluation.

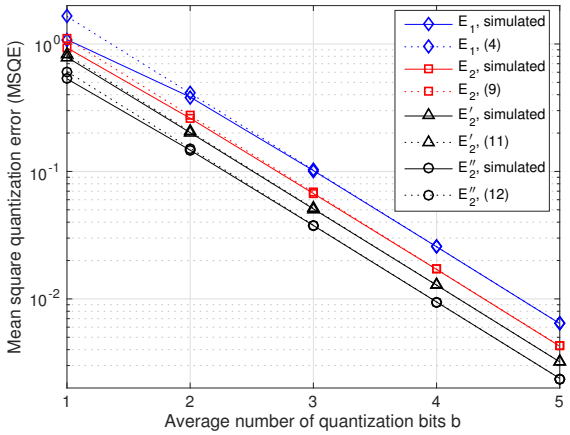


Fig. 3. Mean square quantization error as a function of b, the average number of quantization bits.

Example 1. In this example, we compare the mean square quantization errors when a scalar is quantized using the two structures discussed in Sec. III. Consider the quantization of a random variable $c = re^{j\alpha}$, where r is uniformly distributed over the interval $[0, 2)$ and α is uniformly distributed over $[0, 2\pi)$. The quantization error is averaged over 10^5 realizations. Fig. 3 shows the MSQE for Structure I (S1) and the three quantization schemes of Structure II (S2) discussed in Sec. III. The theoretical errors are computed using the formulas in (4), (9), (11), and (12). Although the quantization error is analyzed in Sec. III under the assumption of high quantization resolution, we can see from Fig. 3 the results are accurate for $b \ge 2$. The theoretical and simulated curves overlap except for the case of very low rate quantization $b = 1$. The simulated gain of S2 over S1 is therefore about the same as the computed gain given in Example 1. In the following examples, joint scalar quantization with uniform quantization of the magnitude will be used for S2.

Example 2. In this example we consider a beamforming system with one RF chain over a channel described by (23). Fig. 4 shows the bit error rate (BER) performance when QPSK symbols are transmitted and $N_t = N_r = 16$. The coefficients of the optimal beamforming and combining vectors are quantized as discussed in Sec. IV-A. The angles ϕ_k and θ_k ((16) for S1 and (17) for S2) are quantized with $b = 1, 1.5$, and 2 bits, where b is the average number of quantization bits. Compared to unconstrained optimal beamforming, the SNR degradation for $b = 2$ is around 0.8 dB with S2 and 2.5 dB with S1. When $b = 1.5$, the angles θ_k and ϕ_k are quantized using 2 bits and one bit, respectively. The performance of S2 with $b = 1.5$ is about the same as that of S1 with $b = 2$ bits; a simpler one-bit quantizer can be used for ϕ_k in S2 to achieve the performance of S1. The gap between the two structures widens as b decreases and the phase shifters are more heavily quantized.

Example 3. Consider a beamforming system with $N_t = N_r = 16$ and one RF chain. Fig. 5 shows the simulated

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2646659, IEEE Transactions on Signal Processing

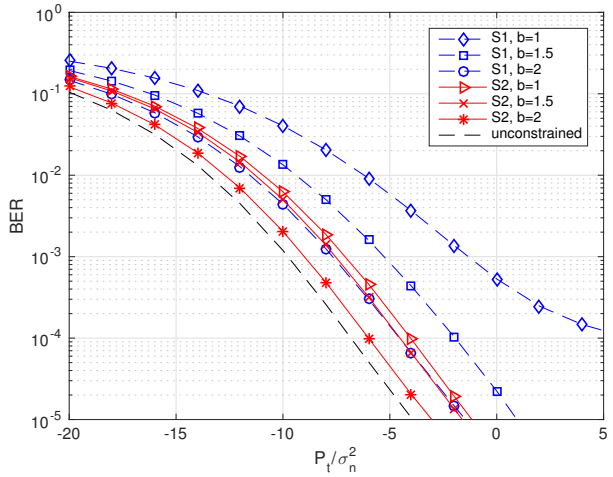


Fig. 4. BER performance of a beamforming system with $N_t = N_r = 16$.

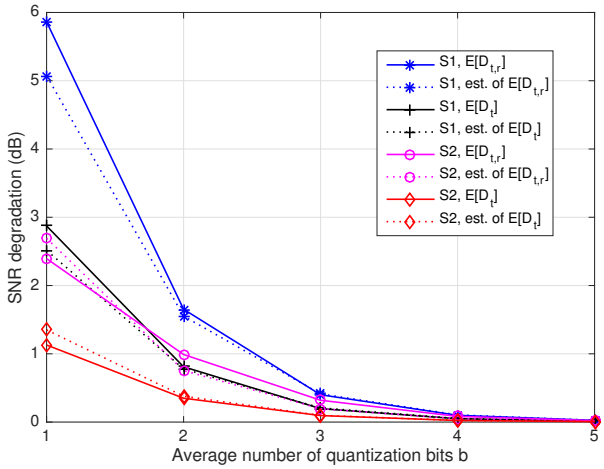
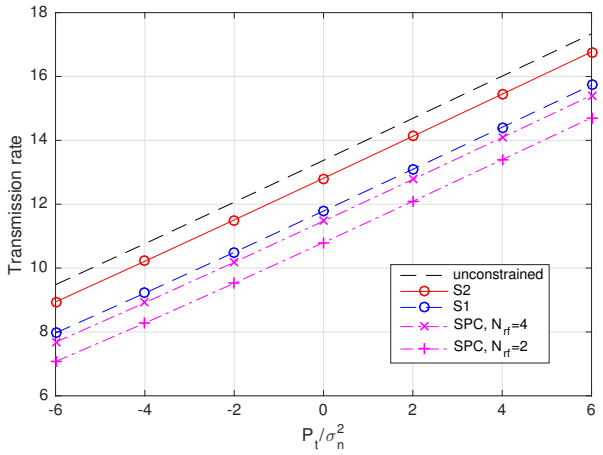


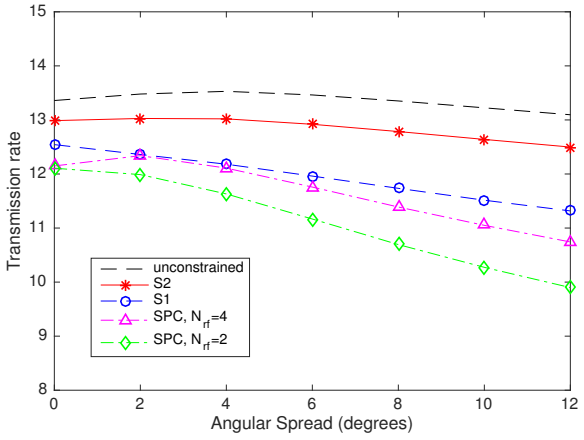
Fig. 5. Average SNR degradation as a function of b for a beamforming system with $N_t = N_r = 16$.

average SNR degradation and the estimated degradation $\overline{\mathcal{D}}_r$ and $\overline{\mathcal{D}}_{t,r}$ computed using (21) and (22). By averaging over 10^5 channels we obtain the average degradation $E[D_t]$ when only the beamformer is quantized and the average degradation $E[D_{t,r}]$ when both the beamformer and combiner are quantized. The crude approximations in (21) and (22) that compute the degradation using the quantization error variance give good estimates of $E[D_t]$ and $E[D_{t,r}]$. Notice that the lower bounds given in (21) and (22) are approximated. They can be larger than the actual SNR degradation for a small b, where approximations are less accurate, as can be observed from Fig. 5.

Example 4. Let $N_t = N_r = 64$, the number of RF chains be $N_{rf} = 2$ and two substreams are transmitted. Fig. 6 shows the transmission rate computed using (1) for $\Gamma = 6.6$, which corresponds to a BER of around 10^{-5}. Equal power allocation is used between the two substreams. The coefficients of $\mathbf{F}_{rf}$ and $\mathbf{G}_{rf}$ are each quantized by quantizing the phase shifters using $b = 2$ as discussed in Sec. V. Also shown in Fig. 6(a) are the curves of the sparse precoding and combining (SPC) [3], in which each coefficient of $\mathbf{F}_{rf}$ is implemented using one phase shifter. The phase shifters in SPC are also quantized using two bits. For comparison we have shown the result of SPC with four RF chains, in which case SPC uses the same total number of phase shifters but twice the number of RF chains. We see that S1 with two RF chains has a higher rate than SPC with four RF chains. To have a comparable performance with that of S1, SPC requires more than twice the number of RF chains and thus more phase shifters. In other words, with S1 (or S2), the number of RF chains can be reduced by half. In 6(b), we plot the transmission rate for different angular spread when $P_t/\sigma_n^2 = 0$ dB. The azimuth and elevation angular spreads at both the transmitter and the receiver are the same. We can see that the performance of S2 quantization is robust as angular spread increases.



(a)



(b)

Fig. 6. Hybrid precoding system with $N_t = N_r = 64$ and $N_s = 2$; (a) transmission rate vs. SNR, and (b) transmission rate vs. angular spread.

Example 5. In the previous examples, point-to-point communication is considered. In this example, we demonstrate

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2646659, IEEE Transactions on Signal Processing

the average rate performance using S1 or S2 for downlink multiuser transmission when $N_t = 32$, $N_r = 8$. The number of RF chains at the base station is the same as the number of users K. The multiuser beamforming vectors are designed based on the zero-forcing beamforming method [24], in which a hybrid beamforming scheme is used at the base station. The analog beamformers of the users and base station are computed without taking interference into consideration and they are quantized using S1 or S2. The digital precoder at the base station is designed so that multiuser interference is canceled and the column vectors of the digital precoder are normalized so that each user is allocated the same power. We show in Fig. 7 the average transmission rate per user as a function of K for $P_t/\sigma_n^2 = 5$ dB. We see that the average rate decreases with the number of users. The gap between the unconstrained and constrained cases increases with K. For $K = 1$, the average rate of S2 with $b = 2$ is around 0.3 bits less than that of the unquantized, but the difference increases to around 0.8 bits for $K = 16$. If the average number of quantization bits b for S2 is increased to 3, the gap can be narrowed to around 0.3 bits for $K = 16$. For more users, the performance is more sensitive to quantization error and a higher quantization resolution is needed to achieve near-optimal performance.

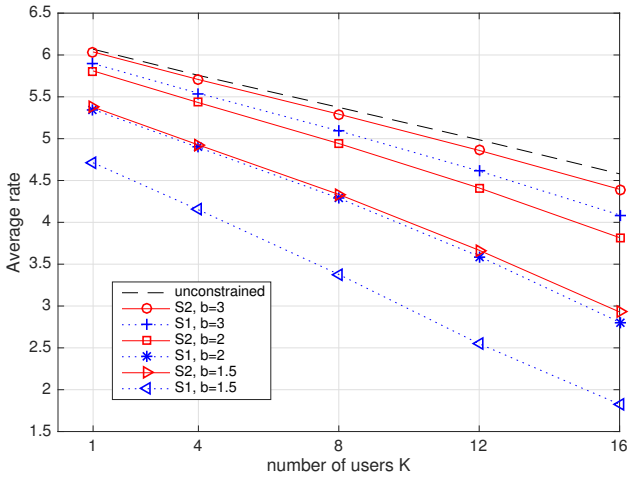


Fig. 7. Average rate per user vs. numbers of users K.

VII. Conclusions

In earlier works of analog RF precoding, the coefficients are optimized with the condition that the coefficients are of constant magnitude. The condition can be much more relaxed. That is, the analog processing consists of only phase shifters and combiners and the coefficients need not be of the same magnitude. The relaxation of the condition gives rise to more general RF precoders and also different possible implementation structures. In this paper we analyze the quantization error of the precoder coefficients due to phase quantization for two possible structures. In these two structures, each coefficient of the RF precoder is represented using two phase shifters, which seemingly increases the number of phase shifters. However simulations show that these structures can reduce the number of RF chains for a given total number of phase shifters.

Appendix A

Proof of Lemma 1. The quantized value $\hat{c} = e^{(j\theta+j\delta_\theta)} + e^{(j\phi+j\delta_\phi)}$ can be written as four terms: $\hat{c} = \cos(\theta+\delta_\theta) + j\sin(\theta+\delta_\theta) + \cos(\phi+\delta_\phi) + j\sin(\phi+\delta_\phi)$. The first term $\cos(\theta+\delta_\theta)$ can be expressed as $\cos(\theta)\cos(\delta_\theta) - \sin(\theta)\sin(\delta_\theta)$. When the quantization resolution is high, δ_θ and δ_ϕ are small and thus $\cos(\delta_\theta) \approx 1$ and $\sin(\delta_\theta) \approx \delta_\theta$. So the first term $\cos(\theta+\delta_\theta)$ can be approximated as $\cos\theta - \delta_\theta \sin\theta$. Applying similar approximations on the remains three terms of $\hat{c}$, we obtain $\hat{c} \approx \cos\theta - \delta_\theta\sin\theta + j(\sin\theta + \delta_\theta\cos\theta) + \cos\phi - \delta_\phi\sin\phi + j(\sin\phi + \delta_\phi\cos\phi)$. The quantization error is $\hat{c} - c \approx -\delta_\theta\sin\theta - \delta_\phi\sin\phi + j(\delta_\theta\cos\theta + \delta_\phi\cos\phi)$. With this expression, we obtain the approximation $\mathcal{E}_1 \approx E[\delta_\theta^2 + \delta_\phi^2 + 2\delta_\theta\delta_\phi\cos(\theta-\phi)]$. Using the assumptions that δ_θ and δ_ϕ are uncorrelated and have zero mean, we get $\mathcal{E}_1 \approx \sigma_{\delta_\theta}^2 + \sigma_{\delta_\phi}^2$.

Appendix B

Proof of Lemma 2. The quantization error can be rewritten as $\hat{c} - c = \cos(\theta+\delta_\theta) + j\sin(\theta+\delta_\theta) + \cos(\theta+\phi+\delta_\theta+\delta_\phi) + j\sin(\theta+\phi+\delta_\theta+\delta_\phi) - e^{j\theta} - e^{j(\theta+\phi)}$. Applying angle sum formulas for cosine and sine functions, we can further express the error as $\hat{c} - c = \cos\theta\cos\delta_\theta - \sin\theta\sin\delta_\theta + j\sin\theta\cos\delta_\theta + j\cos\theta\sin\delta_\theta + \cos(\theta+\phi)\cos(\delta_\theta+\delta_\phi) - \sin(\theta+\phi)\sin(\delta_\theta+\delta_\phi) + j\sin(\theta+\phi)\cos(\delta_\theta+\delta_\phi) + j\cos(\theta+\phi)\sin(\delta_\theta+\delta_\phi) - \cos\theta - j\sin\theta - \cos(\theta+\phi) - j\sin(\theta+\phi)$. Similar to the proof of Lemma 1, let us apply approximations for high-resolution quantization, in particular, $\cos\delta_\theta \approx 1$, $\cos(\delta_\theta+\delta_\phi) \approx 1$, $\sin\delta_\theta \approx \delta_\theta$, and $\sin(\delta_\theta+\delta_\phi) \approx \delta_\theta+\delta_p$. With these approximations, we get $\hat{c} - c \approx -\delta_\theta\sin\theta - (\delta_\theta+\delta_\phi)\sin(\theta+\phi) + j\delta_\theta\cos\theta + j(\delta_\theta+\delta_\phi)\cos(\theta+\phi)$. Combining the terms yields $|\hat{c}-c|^2 \approx 2\delta_\theta^2(1+\cos\phi) + \delta_\phi^2 + 2\delta_\theta\delta_\phi(1+\cos\phi)$. Note that $(1+\cos\phi) = 2\cos^2(\phi/2)$, which is equal to $r^2/2$ as $\phi = 2\cos^{-1}(r/2)$. It follows that $|\hat{c}-c|^2 \approx \delta_\theta^2 r^2 + \delta_\phi^2 + \delta_\theta\delta_\phi r^2$. Using the assumptions that r, δ_θ and δ_ϕ are uncorrelated and also that δ_θ and δ_ϕ have zero mean, we arrive at (7).

Appendix C.

Proof of (10). The quantization error can be expressed as $\hat{c}-c = 2e^{j\alpha}(e^{j\delta_\theta}\cos((\phi+\delta_\phi)/2) - \cos(\phi/2))$. Applying the approximations for high-resolution quantization, $\cos\delta_\theta \approx 1$, $\sin\delta_\theta \approx \delta_\theta$, $\cos(\delta_\phi/2) \approx 1$ and $\sin(\delta_\phi/2) \approx \delta_\phi/2$, we get $\hat{c} - c \approx 2e^{j\alpha}(-\delta_\phi/2\sin(\phi/2) + j\delta_\theta\cos(\phi/2) - j\delta_\theta\delta_\phi/2\sin(\phi/2))$. Notice that the last term contains the product of the two quantization errors, thus it is significantly smaller than the first two. Ignoring the last term, we get $|\hat{c}-c|^2 \approx \delta_\phi^2\sin^2(\phi/2) + 4\delta_\theta^2\cos^2(\phi/2)$. Using $\cos(\phi/2) = r/2$, we obtain the expression in (10).

Appendix D

Proof of (12). Let $\hat{c} = \hat{r}e^{j\hat{\alpha}}$ and the quantization error be $e = c' - c$. With joint scalar quantization, we know $\hat{\alpha} = \alpha + \delta_\theta$ and thus $\hat{c} = e^{j(\alpha+\delta_\theta)}\hat{r}$. The error term $|e|^2 = (\hat{c}-c)^*(\hat{c}-c)$ can be rewritten as $\hat{r}^2 + r^2 - 2r\hat{r}\cos\delta_\theta$. Expressing $\hat{r}$ as $r+\delta_r$, where δ_r is the quantization error of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2016.2646659, IEEE Transactions on Signal Processing

the magnitude, we get $|e|^2 = 4(r^2 + r\delta_r)\sin^2(\delta_\theta/2) + \delta_r^2$. When δ_θ is small, we have $\sin(\delta_\theta/2) \approx \delta_\theta/2$. It follows that $|e|^2 \approx (r^2 + r\delta_r)\delta_\theta^2 + \delta_r^2$. With the high-resolution quantization assumptions A1 and A2, δ_r has zero mean and is uncorrelated with r. Thus we get $\mathcal{E}_2'' \approx E[r^2]\sigma_{\delta_\theta}^2 + \sigma_{\delta_r}^2$, where we have used the assumption the r and δ_θ are uncorrelated. Replacing $\sigma_{\delta_\theta}^2$ by $(\pi^2/3)2^{-2b_\theta}$ and $\sigma_{\delta_r}^2$ by $(1/3)2^{-2b_\phi}$, we get (12).

APPENDIX E

Proof of Lemma 3. The SNR degradation $\mathcal{D}_t$ can be written in term of $\mathbf{f}$ and $\mathbf{d}_f$ as $\mathcal{D}_t = \lambda_0^2||\mathbf{f}+\mathbf{d}_f||^2/||\mathbf{H}(\mathbf{f}+\mathbf{d}_f)||^2$. The denominator $||\mathbf{H}(\mathbf{f}+\mathbf{d}_f)||^2 = \mathbf{f}^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{f} + \mathbf{f}^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{d}_f + \mathbf{d}_f^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{f} + ||\mathbf{H}\mathbf{d}_f||^2$ can be expressed as $\lambda_0^2||\mathbf{f}||^2 + \lambda_0^2(\mathbf{f}^\dagger\mathbf{d}_f + \mathbf{d}_f^\dagger\mathbf{f}) + ||\mathbf{H}\mathbf{d}_f||^2$ using $\mathbf{H}^\dagger\mathbf{H}\mathbf{f} = \lambda_0^2\mathbf{f}$. Therefore, we have

$$\mathcal{D}_t = \frac{1 + (\mathbf{f}^\dagger\mathbf{d}_f + \mathbf{d}_f^\dagger\mathbf{f})/||\mathbf{f}||^2 + ||\mathbf{d}_f||^2/||\mathbf{f}||^2}{1 + (\mathbf{f}^\dagger\mathbf{d}_f + \mathbf{d}_f^\dagger\mathbf{f})/||\mathbf{f}||^2 + ||\mathbf{H}\mathbf{d}_f||^2/(\lambda_0^2||\mathbf{f}||^2)}, \tag{24}$$

where we have divided the numerator and denominator by $\lambda_0^2||\mathbf{f}||^2$. Notice that $||\mathbf{f}||^2$ is in the order of N_t as the entries of $\mathbf{f}$ are normalized to have maximum magnitude equal to 2. When the quantization resolution is high and N_t is large, the value of $\mathbf{f}^\dagger\mathbf{d}_f/||\mathbf{f}||^2 = \mathbf{v}_0^\dagger\mathbf{d}_f/||\mathbf{f}||$ is typically small as $\mathbf{v}_0$ is of unit norm. So $(\mathbf{f}^\dagger\mathbf{d}_f + \mathbf{d}_f^\dagger\mathbf{f})/||\mathbf{f}||^2$, which is two times the real part of $\mathbf{f}^\dagger\mathbf{d}_f/||\mathbf{f}||^2$, will be a small quantity. On the other hand $||\mathbf{H}\mathbf{d}_f||^2/(\lambda_0^2||\mathbf{f}||^2)$ is equal to $||\mathbf{H}\mathbf{d}_f||^2/||\mathbf{H}\mathbf{f}||^2$. Notice that $||\mathbf{H}\mathbf{d}_f||^2/||\mathbf{d}_f||^2 \leq ||\mathbf{H}\mathbf{f}||^2/||\mathbf{f}||^2$ as $\mathbf{f}$ is an eigen vector corresponding to the largest eigen value of $\mathbf{H}^\dagger\mathbf{H}$. Rearranging the terms, we get $||\mathbf{H}\mathbf{d}_f||^2/||\mathbf{H}\mathbf{f}||^2 \leq ||\mathbf{d}_f||^2/||\mathbf{f}||^2$, which is a small number when the quantization resolution is sufficiently high. Therefore the last two terms in the denominator of (24) is much smaller than unity. Using $1/(1+x) \approx 1-x$ for small x, we have $\mathcal{D}_t \approx (1+(\mathbf{f}^\dagger\mathbf{d}_f+\mathbf{d}_f^\dagger\mathbf{f})/||\mathbf{f}||^2+||\mathbf{d}_f||^2/||\mathbf{f}||^2)(1-(\mathbf{f}^\dagger\mathbf{d}_f+\mathbf{d}_f^\dagger\mathbf{f})/||\mathbf{f}||^2-||\mathbf{H}\mathbf{d}_f||^2/(\lambda_0^2||\mathbf{f}||^2))$. Multiplying out the terms, we get $\mathcal{D}_t \approx 1 + ||\mathbf{d}_f||^2/||\mathbf{f}||^2 - \frac{||\mathbf{H}\mathbf{d}_f||^2}{\lambda_0^2||\mathbf{f}||^2} - B1$, where $B1 = \frac{(\mathbf{f}^\dagger\mathbf{d}_f+\mathbf{d}_f^\dagger\mathbf{f})^2}{||\mathbf{f}||^2}\frac{1}{||\mathbf{f}||^2} - \frac{||\mathbf{d}_f||^2}{||\mathbf{f}||^2}\frac{||\mathbf{H}\mathbf{d}_f||^2}{\lambda_0^2||\mathbf{f}||^2} - \frac{(\mathbf{f}^\dagger\mathbf{d}_f+\mathbf{d}_f^\dagger\mathbf{f})}{||\mathbf{f}||^2}\left(\frac{||\mathbf{d}_f||^2}{||\mathbf{f}||^2} + \frac{||\mathbf{H}\mathbf{d}_f||^2}{\lambda_0^2||\mathbf{f}||^2}\right)$. Notice that $B_1 \approx 0$ as each term in B_1 is a product of numbers much smaller than unity and thus (19) follows.

To consider the degradation when both the transmitter and recover are quantized, observe that the denominator term $|\hat{\mathbf{g}}\mathbf{H}\hat{\mathbf{f}}|^2$ in (18) is equal to $|(\mathbf{g}+\mathbf{d}_g)^\dagger\mathbf{g}|^2/\beta_g^2$ using $\mathbf{g} = \beta_g\mathbf{H}\hat{\mathbf{f}}$. It can be expressed as $(||\mathbf{g}||^4 + ||\mathbf{g}||^2(\mathbf{d}_g^\dagger\mathbf{g}+\mathbf{g}^\dagger\mathbf{d}_g) + |\mathbf{d}_g^\dagger\mathbf{g}|^2)/\beta_g^2$ and thus $\mathcal{D}_{t,r} = \lambda_0^2||\hat{\mathbf{f}}||^2||\hat{\mathbf{g}}||^2\beta_g^2/(||\mathbf{g}||^4 + ||\mathbf{g}||^2(\mathbf{d}_g^\dagger\mathbf{g}+\mathbf{g}^\dagger\mathbf{d}_g) + |\mathbf{d}_g^\dagger\mathbf{g}|^2)$. Dividing the numerator and the denominator by $||\mathbf{g}||^4$, we have

$$\mathcal{D}_{t,r} = \frac{\mathcal{D}_t||\hat{\mathbf{g}}||^2/||\mathbf{g}||^2}{1 + (\mathbf{d}_g^\dagger\mathbf{g}+\mathbf{g}^\dagger\mathbf{d}_g)/||\mathbf{g}||^2 + |\mathbf{d}_g^\dagger\mathbf{g}|^2/||\mathbf{g}||^4},$$

where we have used $\mathbf{g} = \beta_g\mathbf{H}\hat{\mathbf{f}}$ and $\mathcal{D}_t = \frac{\lambda_0^2||\hat{\mathbf{f}}||^2}{||\mathbf{H}\hat{\mathbf{f}}||^2}$. Notice that in the denominator the last term is a lot smaller than the first two due to the factor $1/||\mathbf{g}||^4$ and also that the second term is much smaller than unity. Ignoring the last term in the denominator and using $1/(1+x) \approx 1-x$ for small x, we have $\mathcal{D}_{t,r} \approx \mathcal{D}_t||\hat{\mathbf{g}}||^2/||\mathbf{g}||^2(1-(\mathbf{d}_g^\dagger\mathbf{g}+\mathbf{g}^\dagger\mathbf{d}_g)/||\mathbf{g}||^2)$. As $||\hat{\mathbf{g}}||^2/||\mathbf{g}||^2 = 1 + (\mathbf{d}_g^\dagger\mathbf{g}+\mathbf{g}^\dagger\mathbf{d}_g)/||\mathbf{g}||^2 + ||\mathbf{d}_g||^2/||\mathbf{g}||^2$, we have $\mathcal{D}_{t,r} \approx \mathcal{D}_t\left(1 + ||\mathbf{d}_g||^2/||\mathbf{g}||^2 - B_2\right)$, where $B_2 = (\mathbf{d}_g^\dagger\mathbf{g}+\mathbf{g}^\dagger\mathbf{d}_g)^2/||\mathbf{g}||^4 + (\mathbf{d}_g^\dagger\mathbf{g}+\mathbf{g}^\dagger\mathbf{d}_g)||\mathbf{d}_g||^2/||\mathbf{g}||^4$, a number much smaller than $||\mathbf{d}_g||^2/||\mathbf{g}||^2$. Ignoring B_2, we arrive at (20).

REFERENCES

[1] C. H. Doan, S. Emami, D. A. Sobel, A. M. Niknejad, and R. W. Brodersen, "Design considerations for 60 GHz CMOS radios," *IEEE Communications Magazine,* vol. 42, no. 12, pp. 132-140, Dec. 2004.

[2] Z. Pi and F. Khan, "An introduction to millimeter-wave mobile broadband systems," *IEEE Communications Magazine*, vol. 49, no. 6, pp. 101-107, June 2011.

[3] O. E. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath Jr, "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 99, pp. 1-15, Jan. 2014.

[4] S. Kutty and D. Sen, "Beamforming for millimeter wave communications: an inclusive survey," *IEEE Communications Surveys and Tutorials*, Dec. 2015.

[5] X. Zhang, A. F. Molisch, and S. Y. Kung, "Variable-phase-shift-based RF-baseband codesign for MIMO antenna selection," *IEEE Trans. Signal Processing*, vol. 53, no. 11, pp. 4091-4103, Nov. 2005.

[6] E. Zhang and C. Huang, "On achieving optimal rate of digital precoder by RF-baseband codesign for MIMO systems," Proceedings of IEEE Vehicular Technology Conference, Sept. 2014

[7] T. E. Bogale, L. B. Le, A. Haghighat, and L. Vandendorpe, "On the number of RF chains and phase shifters, and scheduling design with hybrid analog-digital beamforming," 2014, http://arxiv.org/abs/1410.2609.

[8] A. Alkhateeb, J. Mo, N. G. Prelcic, and R. W. Heath Jr, "MIMO precoding and combining solutions for millimeter-wave systems," *IEEE Communications Magazine*, vol.52, no.12, pp.122-131, Dec. 2014.

[9] J. A. Zhang, X. Huang, V. Dyadyuk, and Y. J. Guo, "Massive hybrid antenna array for millimeter-wave cellular communications," *IEEE Wireless Communications*, pp. 79-87, Feb. 2015.

[10] J. D. Krieger, C.-P. Yeang, and G. W. Wornell, "Dense delta-sigma phased arrays," *IEEE Transactions on Antennas and Propagation*, vol. 61, no. 4, April 2013.

[11] S. Payami, M. Shariat, M. Ghoraishi, and M. Dianati, "Effective RF codebook design and channel estimation for millimeter wave communication systems", IEEE International Conference on Communication Workshop, June 2015.

[12] F. Sohrabi and W. Yu, "Hybrid beamforming with finite-resolution phase shifters for large-scale MIMO systems," IEEE International Workshop on Signal Processing Advances in Wireless Communications, 2015

[13] N. Benvenuto, S. Montagner, and L. Pilloni, "Vector quantized phase shift analog beamformers for millimeter-wave systems," IEEE Latin-America Conference on Communications, 2014

[14] S. Chang, W. Hong, and J. Oh, "Quantization effects of phase shifters on 5G mmWave antenna arrays," IEEE International Symposium on Antennas and Propagation, 2015.

[15] A. J. Goldsmith and S.-G. Chua, "Variable-rate variable-power MQAM for fading channels," *IEEE Trans. Communi.*, vol. 45, no. 10, pp. 1218-1230, Oct. 1997.

[16] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Kluwer Academic Publishers, 1991.

[17] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*, Prentice-Hall, 1993.

[18] T. E. Bogale, L. B. Le, A. Haghighat, and L. Vandendorpe, "On the number of RF chains and phase shifters, and scheduling design with

hybrid analog-digital beamforming," IEEE Trans. Wireless Communications, vol. 15, no. 5, pp. 3311-3326, May 2016.

[19] A. Alkhateeb, Y. H. Nam, J. Zhang, and R. W. Heath Jr, "Massive MIMO combining with switches," *IEEE Communications Letters*, vol.5, no. 3, pp.232-235, June 2016.

[20] A. Alkhateeb and R. W. Heath Jr, "Frequency Selective Hybrid Precoding for Limited Feedback Millimeter Wave Systems," IEEE Transactions on Communications, vol. 64, no. 5, pp.1801-1818, May 2016.

[21] V. Erceg, et al., "TGn channel models," IEEE 802.11-03/940r4, May 2004.

[22] M. R. Akdeniz, Y. Liu, M. K. Samimi, S. Sun, S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter Wave Channel Modeling and Cellular Capacity Evaluation," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1164 - 1179, June 2014.

[23] K. I. Pedersen, P. E. Mogensen, and B. H. Fleury, "The Power azimuth spectrum in outdoor environments," *IEE Electronics Letters*, vol. 33, no. 18, pp. 1583-1584, Aug. 1997.

[24] Ahmed Alkhateeb, Geert Leus, and R. W. Heath Jr, "Limited Feedback Hybrid Precoding for Multi-User Millimeter Wave Systems," *IEEE Transactions on Wireless Communications*, vol. 14, no. 11, pp. 6481–6494, Nov. 2015.

Yuan-Pei Lin Biosketch. Yuan-Pei Lin (S'93-M'97-SM'03) was born in Taipei, Taiwan, 1970. She received the B.S. degree in control engineering from the National Chiao-Tung University, Taiwan, in 1992, and the M.S. degree and the Ph.D. degree, both in electrical engineering from California Institute of Technology, in 1993 and 1997, respectively. She joined the Department of Electrical and Control Engineering of National Chiao-Tung University, Taiwan, in 1997. Her research interests include digital signal processing, multirate filter banks, and signal processing for digital communications.

She was a recipient of Ta-You Wu Memorial Award in 2004. She served as an associate editor for IEEE Transaction on Signal Processing, IEEE Transaction on Circuits and Systems II, IEEE Signal Processing Letters, IEEE Transaction on Circuits and Systems I, EURASIP Journal on Applied Signal Processing, and Multidimensional Systems and Signal Processing, Academic Press. She was a distinguished lecturer of the IEEE Circuits and Systems Society for 2006-2007. She has also coauthored two books, Signal Processing and Optimization for Transceiver Systems, and Filter Bank Transceivers for OFDM and DMT Systems, both by Cambridge University Press, 2010