FEEDBACK REDUCTION FOR SPATIAL MULTIPLEXING WITH LINEAR PRECODING

*Claude Simon and Geert Leus**

Delft University of Technology, Fac. EEMCS, Mekelweg 4, 2628 CD Delft, The Netherlands

ABSTRACT

This paper presents two novel methods to optimally compress the feedback for spatial multiplexing with linear precoding. The methods exploit the time correlation of the channel and the knowledge of the previously fed back precoder matrices to estimate the conditional probabilities of the different possible feedback indices. These probabilities are then used to losslessly compress the actual feedback using variable-length codes. Two compression schemes are presented, one for a non-dedicated feedback channel and one for a dedicated feedback channel.

***Index Terms*—** MIMO systems, linear precoding, partial CSI feedback

1. INTRODUCTION

In the last few years spatial multiplexing emerged as a promising scheme to fulfill the data rate requirements of future wireless services. A technique to make spatial multiplexing more robust to rank deficient channels and to allow for simpler receiver architectures is linear precoding [1].

The optimal precoder matrix is calculated as a function of the channel state information (CSI). However, since CSI is in general just available at the receiver, it requires to be fed back to the transmitter. Since the precoder matrix is generally restricted to be unitary it is beneficial [2] to feed back the quantized precoder matrix, instead of the quantized channel matrix.

The feedback requirements can be further reduced by exploiting the temporal correlation of the channel. In [3], a first-order Markov chain is introduced to model the feedback of a beamforming vector. Based on this Markov model, no feedback is sent if the current state is the same as the previous state, whereas a fixed-length code is fed back for all other states. In [4], this approach is extended, by ignoring the states with low probability, thereby reducing the length of the fixed-length code and thus reducing the feedback requirements. Both methods are suboptimal though, since they either assign no code or a fixed-length code to a state. Moreover, whereas the compression adopted in [3] is lossless, the compression in [4] is lossy. Further, modeling the beamforming vector as a first-order Markov chain assumes that the current beamforming vector just depends on the previous one. This assumption discards the information from the previously fed back beamforming vectors, which may be exploited to improve the prediction.

In this paper, we model the feedback of a linear precoder (instead of a beamforming vector) using a higher-order Markov chain (instead of a first-order Markov chain). Moreover, we use techniques from optimal source coding to reduce the average feedback rate. Compared to existing suboptimal feedback-reduction schemes, the encoding is variable-length and lossless.

*This research was supported in part by NWO-STW under the VIDI program (DTC.6577).

Notation: Vectors are designated with lowercase boldface letters, and matrices with capital boldface letters. The notation $[\mathbf{A}]_{i,j}$ denotes the (i,j)th entry of the matrix $\mathbf{A}$. $\mathbf{I}_n$ is the $n \times n$ identity matrix. Further, $\mathbf{A}^H$ denotes the conjugate transpose of the matrix $\mathbf{A}$, and $\mathbf{A}^{-1}$ the inverse. The cardinal number of the set $\mathcal{A}$ is denoted $|\mathcal{A}|$. In addition, $\text{abs}(\mathbf{A})$ represents the element-wise absolute value, $\text{diag}(\mathbf{A})$ a diagonal matrix obtained by removing the off-diagonal elements of $\mathbf{A}$, and $\|\mathbf{A}\|$ the Frobenius norm of $\mathbf{A}$. Finally, $E(\cdot)$ represents expectation, and $P(\cdot)$ probability.

2. SYSTEM MODEL

We assume a narrowband spatial multiplexing MIMO system, with N_T transmit and N_R receive antennas. The system transmits $N_S \leq \min(N_T, N_R)$ symbol streams. The input-output relation, at time instant n, is

$$\mathbf{y}[n] = \mathbf{H}[n]\mathbf{F}[n]\mathbf{s}[n] + \boldsymbol{\nu}[n], \tag{1}$$

where $\mathbf{y}[n] \in \mathbb{C}^{N_R \times 1}$ is the received vector, $\mathbf{s}[n] \in \mathbb{C}^{N_S \times 1}$ is the data symbol vector, $\mathbf{H}[n] \in \mathbb{C}^{N_R \times N_T}$ is the channel matrix, $\mathbf{F}[n] \in \mathbb{C}^{N_T \times N_S}$ is the linear precoder matrix, and $\boldsymbol{\nu}[n] \in \mathbb{C}^{N_R \times 1}$ is the additive noise vector. We assume that the elements of $\mathbf{s}[n]$ are i.i.d. and uniformly distributed over a finite alphabet $\mathcal{A}$ with zero mean and variance 1. We further assume that the elements of $\boldsymbol{\nu}[n]$ are i.i.d. and complex Gaussian distributed with zero mean and variance 1. We finally assume that the elements of $\mathbf{H}[n]$ are i.i.d. and complex Gaussian distributed with zero mean and variance P, and that every element is distributed in time according to Jakes' model [5] with Doppler frequency f_D. The singular value decomposition (SVD) of $\mathbf{H}[n]$ will be denoted as $\mathbf{H}[n] = \mathbf{U}[n]\boldsymbol{\Sigma}[n]\mathbf{V}^H[n]$, where $\mathbf{U}[n]$ and $\mathbf{V}[n]$ belong to $\mathcal{U}_{N_R \times N_R}$ and $\mathcal{U}_{N_T \times N_T}$, respectively, with $\mathcal{U}_{n \times m}$ denoting the set of unitary $n \times m$ matrices, and $\boldsymbol{\Sigma}[n]$ is a diagonal $N_R \times N_T$ matrix with the diagonal starting in the top left corner. For later use, we will also define $\bar{\mathbf{U}}[n] = [\mathbf{U}[n]]_{:,1:N_S} \in \mathcal{U}_{N_R \times N_S}$, $\bar{\mathbf{V}}[n] = [\mathbf{V}[n]]_{:,1:N_S} \in \mathcal{U}_{N_T \times N_S}$, and $\bar{\boldsymbol{\Sigma}}[n] = [\boldsymbol{\Sigma}[n]]_{1:N_S,1:N_S}$. The transmission occurs blockwise, every block consists of N symbol vectors, and we assume perfect CSI at the receiver at the beginning of each block. Further, the feedback link to the transmitter is assumed to be delay-free and error-free, but bandwidth limited.

Precoder codebooks, and selection criteria to pick an entry from the codebook are discussed in the next section. The subsequent section then explains how the selected precoder index is compressed before it is fed back to the transmitter.

3. PRECODER CODEBOOKS AND SELECTION

Assuming that the data rate on the feedback channel is limited, we need to quantize the precoder matrix. Quantization requires a codebook, which contains the quantized precoder matrices, and a selection criterion, which maps the estimated channel to an entry of the codebook. Both of course strongly depend on each other.

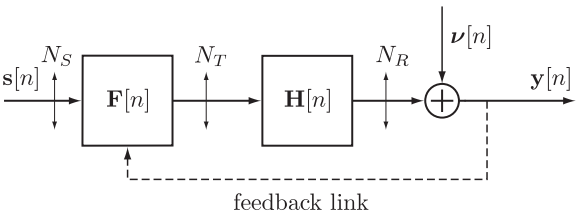

Fig. 1. System model

Generally, in order to reduce the feedback [6], the precoder is limited to be unitary, i.e., $\mathbf{F} \in \mathcal{U}_{N_T \times N_S}$. For most performance criteria, the optimal unitary precoder is given by $\mathbf{F} = \bar{\mathbf{V}}$ [6]. For the BER, however, the optimal precoder is still unknown [7]. However, for realistic SNR's it has been shown in [7] that the optimal unitary precoder is given by $\mathbf{F} = \bar{\mathbf{V}}\mathbf{M}$, where $\mathbf{M} \in \mathcal{U}_{N_S \times N_S}$ with constant modulus entries, e.g., the Hadamard or the DFT matrix.

3.1. Codebook Design

Presently, all the existing codebooks for linear precoding are calculated through iterative algorithms [8, 9], since no analytical solution for the corresponding design criteria exists. All these algorithms select the codebook entries $\mathbf{F}_i$ and the related channel regions $\mathcal{R}_i$ so that the expectation of a distortion function between the channel and the quantized precoders is minimized

$$\{\mathcal{R}_i, \mathbf{F}_i\} = \underset{\{\mathcal{R}_i, \mathbf{F}_i\} | \mathcal{R}_i \subset \mathbb{C}^{N_R \times N_T}, \mathbf{F}_i \in \mathcal{U}_{N_T \times N_S}}{\arg\min} \sum_i E(d(\mathbf{H}, \mathbf{F}_i | \mathbf{H} \in \mathcal{R}_i)) P(\mathbf{H} \in \mathcal{R}_i). \quad (2)$$

In [6], a number of performance criteria has been transformed into subspace distances between $\bar{\mathbf{V}}$ and $\mathbf{F}_i$, such as the chordal distance, the projection two-norm, and the Fubini-Study distance. The squares of these subspace distances are then used in (2) as a distortion measure. Note that $d(\mathbf{H}, \mathbf{F}_i)$ then is actually independent of the signal-to-noise ratio (SNR), since the only knowledge required about $\mathbf{H}$ is $\bar{\mathbf{V}}$. In [6], (2) is not solved through the Lloyd algorithm [8], but the problem is transformed into a subspace packing problem on a Grassmann manifold and algorithms presented in [9] are adopted. In [7], however, the Lloyd algorithm is adopted to solve (2), leading to slightly improved codebooks.

The distortion measure adopted in [10] is related to the capacity loss introduced by quantization, and in contrast to the subspace distances mentioned earlier, it depends on the SNR:

$$d_L(\mathbf{H}, \mathbf{F}_i) = \mathrm{tr}(\hat{\bar{\mathbf{\Sigma}}}^2 - \hat{\bar{\mathbf{\Sigma}}}^2 \bar{\mathbf{V}}^H \mathbf{F}_i \mathbf{F}_i^H \bar{\mathbf{V}}), \quad (3)$$

where $\hat{\bar{\mathbf{\Sigma}}}^2 = (\mathbf{I}_{N_S} + \bar{\mathbf{\Sigma}}^2)^{-1} \bar{\mathbf{\Sigma}}^2$. Note the close resemblance to the squared chordal distance:

$$d_c(\mathbf{H}, \mathbf{F}_i) = \mathrm{tr}(\mathbf{I}_{N_S} - \bar{\mathbf{V}}^H \mathbf{F}_i \mathbf{F}_i^H \bar{\mathbf{V}}), \quad (4)$$

which is independent of the SNR.

A common problem to all the above distortion measures is that all the precoder matrices in the same subspace have the same distortion, despite having different BER performances. In order to solve this problem, one could for instance encode $\bar{\mathbf{V}}$ by retaining the order of the modes, which has the potential of leading to a better BER performance, when we replace the selected $\mathbf{F}_i$ by $\mathbf{F}_i\mathbf{M}$ [7]. In [11], for instance, the squared Frobenius norm distance between $\bar{\mathbf{V}}$ and $\mathbf{F}_i$ was considered:

$$\begin{aligned} d_F(\mathbf{H}, \mathbf{F}_i) &= \|\bar{\mathbf{V}} - \mathbf{F}_i\|^2 \\ &= 2\,\mathrm{tr}[\mathbf{I}_{N_S} - \Re(\bar{\mathbf{V}}^H \mathbf{F}_i)]. \end{aligned} \quad (5)$$

This approach was modified in [12] to take the phase ambiguity of the right singular vectors into account:

$$\begin{aligned} d_{mF}(\mathbf{H}, \mathbf{F}_i) &= \|\bar{\mathbf{V}}\mathrm{diag}(\bar{\mathbf{V}}^H \mathbf{F}_i)\mathrm{diag}^{-1}(\mathrm{abs}(\bar{\mathbf{V}}^H \mathbf{F}_i)) - \mathbf{F}_i\|^2 \\ &= 2\,\mathrm{tr}[\mathbf{I}_{N_S} - \mathrm{abs}(\bar{\mathbf{V}}^H \mathbf{F}_i)]. \end{aligned} \quad (6)$$

Note that these two measures are again independent of the SNR. Also observe the difference with the squared chordal distance of (4). The problem with the two above distortion measures is that the centroid computation required for the Lloyd algorithm is difficult to carry out in closed form. Hence, we apply a brute-force centroid computation by exhaustively searching for the best center, i.e., the channel which has the minimal average distortion within a region. This approach could actually also be used for a BER distortion measure, which is currently under investigation.

3.2. Selection Criteria

In [6], many different selection criteria have been proposed depending on the chosen codebook. In other words, if the codebook is derived based on a certain performance measure, such as for instance symbol MSE of the linear MMSE receiver, this performance measure is also used as a selection criterion.

In [7], however, a BER selection criterion is used, which selects the codeword $\mathbf{F}_i$ from the codebook $\mathcal{F}$ that minimizes the BER of the transmission:

$$\mathbf{F} = \underset{\mathbf{F}_i \in \mathcal{F}}{\arg\min}\, \mathrm{BER}(\mathbf{H}, \mathbf{F}_i) \quad (7)$$

where $\mathrm{BER}(\mathbf{H}, \mathbf{F})$ represents the average BER for the channel $\mathbf{H}$ using the precoder $\mathbf{F}$, and any possible receiver. For a linear receiver, the BER can be calculated with the help of the exact BER expressions [13] for an AWGN channel, using a square or rectangular constellation. This is of course the ultimate selection criterion, which we will adopt in our simulations. Note, however, that in case we use d_F or d_{mF} as distortion measure, we replace the selected $\mathbf{F}_i$ by $\mathbf{F}_i\mathbf{M}$ [7].

4. FEEDBACK REDUCTION

The following section explores two new methods to reduce the feedback requirements for linear precoding with spatial multiplexing. The two methods assume a different underlying system model. The idea behind both models is that the feedback index is encoded before it is sent over the feedback link.

The first model assumes a non-dedicated feedback channel, i.e., the feedback link is also used for data transmission. Thus, the transmitter needs to know when the codeword ends and when the data transmission starts.

The second model, on the other hand, assumes that the feedback link is only used to feed back the index of the precoder matrices. The advantage of reducing the average feedback length for this model is that less energy is required for the feedback.

However, the general feedback algorithm for both schemes is the same, both models just differ in the source encoding that is used.

4.1. Feedback Encoding

The general algorithm for both models is identical. At the start of each block we assume perfect CSI available at the receiver. The optimal precoder is calculated based on the CSI, and then, according to the used selection criterion, it is quantized to the closest entry in the codebook. The index of the selected codebook entry is then encoded, depending on the system model, i.e., if it is a non-dedicated or dedicated feedback channel. In order to exploit the time correlation of the channel we use the transition probability of every codebook entry, i.e., the conditional probability of every codebook entry given the previous K codebook entries. Note that the transition probabilities do not just depend on the past channel states alone, but also on the channel characteristics. Similar to [3], where $N_S = 1$ and $K = 1$ is considered, we rely on Monte-Carlo methods to determine the transition probabilities

$$P_{\mathbf{F}_i,\mathbf{F}_{j_1},\ldots,\mathbf{F}_{j_K}} = P\left(\mathbf{F}[kN] = \mathbf{F}_i | \mathbf{F}[(k-1)N] = \mathbf{F}_{j_1}, \ldots , \mathbf{F}[(k-K)N] = \mathbf{F}_{j_K}\right). \quad (8)$$

Assuming that the channel characteristics do not change, the transition probabilities just depend on the last K quantized precoders. Every set of K quantized precoders corresponds to a different set of transition probabilities and thus to a different compression scheme. For small numbers of K the probabilities, and thus the compression schemes, can be calculated offline. The number of compression schemes that have to be stored at the transmitter and receiver is $N_C = |\mathcal{F}|^K$.

The receiver knows the last K quantized precoders, and encodes the current quantized precoder based on the related compression scheme. The encoded feedback index is then sent to the transmitter. Since the transmitter also knows the last K quantized precoders, it also knows which compression scheme was used and can decode the feedback index.

4.2. Non-Dedicated Feedback Channel

A non-dedicated channel requires the indices to be instantaneously decodable, i.e., they have to satisfy the prefix condition [14]: a codeword can not contain any other codeword as a prefix. However, designing a prefix-free code with a small average length depends on the correct determination of the transition probabilities of the different codebook entries. As a prefix-free code we select the Huffman code. The necessary transition probabilities are estimated through Monte-Carlo simulations. An example of such a compression scheme is depicted in Table 1.

4.3. Dedicated Feedback Channel

The advantage of a dedicated feedback channel is that non-prefix-free (NPF) codes, i.e., codes which do not satisfy the prefix condition, can be used. This is possible because the start and the end of the codeword are easily determined. These codes do not require the knowledge of the exact transition probabilities, but just need the order of the transition probabilities. Further, we are not required to apply feedback for every block, e.g., no feedback means that the actual precoder is the most probable codebook entry. Hence, we do not assign any codeword to the the most probable precoder, and we gradually assign longer and longer codewords to the other precoders in decreasing order of probability. See again Table 1 for an example.

A dedicated channel was also assumed in [3, 4]. However, in those schemes either no codeword or a fixed-length codeword was sent, in a lossless [3] or lossy [4] fashion. These codes clearly are sub-optimal compared to the proposed NPF code.

Codebook	$P_{\mathbf{F}_i,\mathbf{F}_8}$	Huffman Code	NPF Code
$\mathbf{F}_8$	0.25	01	/
$\mathbf{F}_2$	0.20	11	0
$\mathbf{F}_7$	0.18	000	1
$\mathbf{F}_4$	0.16	001	00
$\mathbf{F}_3$	0.10	101	01
$\mathbf{F}_6$	0.08	1000	10
$\mathbf{F}_5$	0.02	10010	11
$\mathbf{F}_1$	0.01	10011	000

Table 1. Example of a compression scheme for $K = 1$.

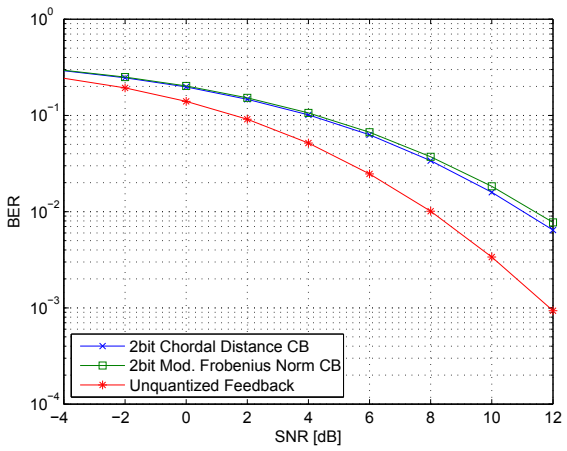

Fig. 2. BER performance for different codebooks (2 bits) and the BER selection criterion. ZF detector, 4-QAM, $N_S = 2$, $N_T = 4$, and $N_R = 2$.

5. SIMULATIONS

We consider a spatial multiplexing MIMO system which transmits 2 symbol streams ($N_S = 2$) over 4 transmit antennas ($N_T = 4$), and has 2 receive antennas ($N_R = 2$). The system uses a ZF receiver, and 4-QAM is adopted. A comparison of two codebooks using the BER selection strategy is depicted in Figs. 2 and 3 for codebooks with 4 entries (2 bits) and 64 entries (6 bits), respectively. One codebook is based on the squared chordal distance d_c, whereas the other is based on the modified Frobenius norm d_{mF}. Although, the second codebook was expected to have a better BER than the first codebook, because optimal mixing of the modes with the matrix $\mathbf{M}$ can be carried out, this effect seems to show up only for large codebooks.

The performance of the two presented feedback compression schemes is depicted in Fig. 4. We see that the NPF code performs better than the Huffman code. The Doppler spread in the simulation is fixed to $f_D = 30$ Hz, and the block length is varied. For very small block lengths, the channel does not change much in between feedback instances, thus the exploitation of the time correlation reduces efficiently the average feedback length. The Huffman code approaches 1 bit feedback, and the NPF code 0 bit feedback. The Huffman code assigns in this case a single bit to the most probable codeword, and the non-prefix-free code does not perform any feedback. If the channel is changing very fast, all the precoders become

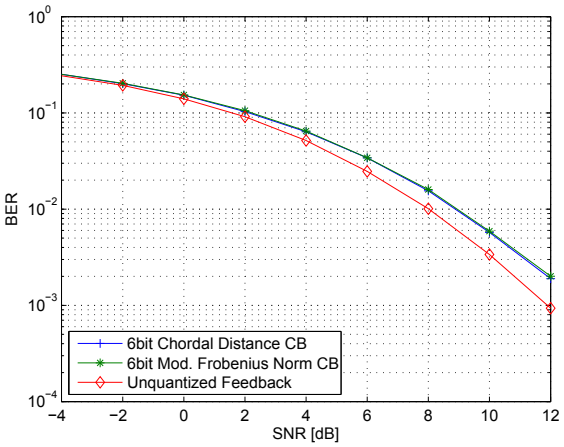

Fig. 3. BER performance for different codebooks (6 bits) and the BER selection criterion. ZF detector, 4-QAM, $N_S = 2$, $N_T = 4$, and $N_R = 2$.

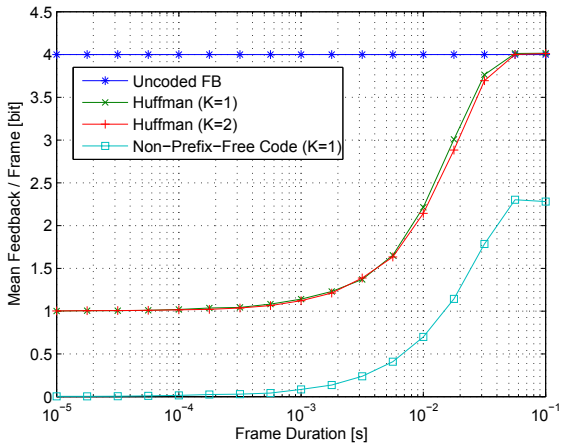

Fig. 4. Feedback reduction for different compression schemes, ZF detector, 4-QAM, $N_T = N_R = 2$, 4 bit codebook, $f_D = 30$ Hz.

equally probable and the Huffman code assigns codewords with the same length for all the indexes ($\lceil \log_2 |\mathcal{F}| \rceil$). The average feedback rate of the NPF code, however, just converges to the average length of the codewords ($1/|\mathcal{F}| \sum_{i=1}^{|\mathcal{F}|} \lfloor \log_2 i \rfloor$).

6. CONCLUSIONS

In this paper we have presented two schemes to reduce the average feedback length through variable-length compression of the precoder index. We considered two different types of compression, and compared their performance through Monte-Carlo simulations. Which of these schemes can be used entirely depends on the assumptions on the feedback link.

7. REFERENCES

[1] Anna Scaglione, Petre Stoica, Sergio Barbarossa, Georgios B. Giannakis, and Hemanth Sampath, "Optimal designs for space-time linear precoders and decoders," *IEEE Trans. Signal Processing*, vol. 50, no. 5, pp. 1051–1064, May 2002.

[2] David J. Love and Robert W. Heath Jr., "Limited feedback precoding for spatial multiplexing systems using linear receivers," in *Proc. Military Communications Conference (MILCOM'03)*, Oct. 2003, vol. 1, pp. 627 – 632.

[3] Kaibin Huang, Bishwarup Mondal, Robert W. Heath Jr., and Jeffrey G. Andrews, "Markov models for limited feedback MIMO systems," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'06)*, May 2006, vol. 4, pp. IV–9–IV–12, unknown.

[4] Kaibin Huang, Bishwarup Mondal, Robert W. Heath Jr., and Jeffrey G. Andrews, "Multi-antenna limited feedback for temporally-correlated channels: Feedback compression," in *Proc. Global Telecommunications Conf. (GLOBECOM'06)*, Dec. 2006.

[5] William C. Jakes, Jr., *Microwave Mobile Communications*, John Wiley & Sons, 1974.

[6] David J. Love and Robert W. Heath Jr., "Limited feedback unitary precoding for spatial multiplexing systems," *IEEE Trans. Inform. Theory*, vol. 51, no. 8, pp. 2967–2976, Aug. 2005.

[7] Shengli Zhou and Baosheng Li, "BER criterion and codebook construction for finite-rate precoded spatial multiplexing with linear receivers," *IEEE Trans. Signal Processing*, vol. 54, no. 5, pp. 1653–1665, May 2006.

[8] Allen Gersho and Robert M. Gray, *Vector Quantization and Signal Compressing*, Kluwer Academic Publishers, 1995.

[9] Bertrand M. Hochwald, Thomas L. Marzetta, Thomas J. Richardson, Wim Sweldens, and Rüdiger Urbanke, "Systematic design of unitary space-time constellations," *IEEE Trans. Inform. Theory*, vol. 46, no. 6, pp. 1962–1973, Sept. 2000.

[10] June Chul Roh and Bhaskar D. Rao, "Design and analysis of MIMO spatial multiplexing systems with quantized feedback," *IEEE Trans. Signal Processing*, vol. 54, no. 8, pp. 2874–2886, Aug. 2006.

[11] Nadia Khaled, Bishwarup Mondal, Robert W. Heath Jr., Geert Leus, and Frederik Petré, "Quantized multi-mode precoding for spatial multiplexing MIMO-OFDM systems," in *Proc. IEEE 62nd Vehicular Technology Conference (VTC'05)*, Sept. 2005, vol. 2, pp. 867–871.

[12] Geert Leus, Claude Simon, and Nadia Khaled, "Spatial multiplexing with linear precoding in time-varying channels with limited feedback," in *Proc. 14th European Signal Processing Conference (EUSIPCO'06)*, Florence, IT, Sept. 2006.

[13] Kyongkuk Cho and Dongweon Yoon, "On the general BER expression of one- and two-dimensional amplitude modulations," *IEEE Trans. Commun.*, vol. 50, no. 7, pp. 1074–1080, July 2002.

[14] John G. Proakis, *Digital Communications*, McGraw-Hill, 2000.