Adaptive transmit beamforming with closed loop feedback in MIMO systems

Eduardo Zacarías B.
Signal Processing Laboratory
ezacaria@wooster.hut.fi

***Abstract*— Adaptive transmit beamforming techniques have been proposed for wireless systems to improve the link level performance. In Time Division Duplexing (TDD) systems, a close to optimal transmit beamforming gain can be reached, since the forward channel can be estimated from the reverse channel measurements. However, in Frequency Division Duplexing (FDD) systems as WCDMA, the Channel Side Information (CSI) at the transmitter must come from the closed loop feedback channel. Current WCDMA systems have limited capacity in this link [1], which restricts the amount of information that can be signaled. With high mobile velocities, the feedback delay plays a key role in performance and, therefore, the adaptive beamforming schemes need to be based on short feedback messages. This document provides an overview of different techniques that exploit the CSI by using adaptive transmit beamforming.**

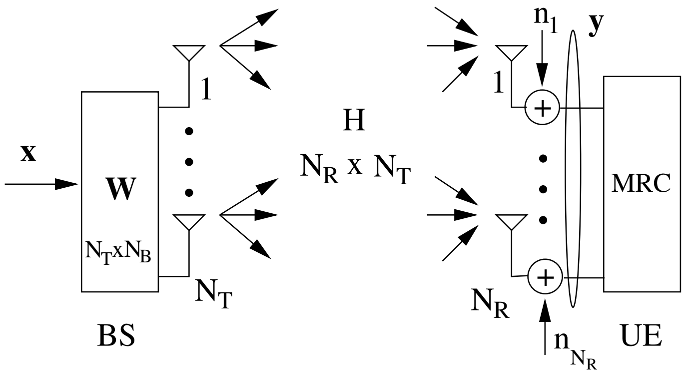

Fig. 1. System model

I. Introduction

The goal of adaptive beamforming is to maximize the instantaneous received power at the UE, by adapting the complex baseband weights of the transmission antennas, based on the closed loop feedback information. For the particular case of two transmit antennas, two schemes (known as "Mode 1" and "Mode 2") have been defined in the WCDMA standard [1]. However, the extension to arrays with more than two antennas is problematic for fast fading channels, since weights for individual antennas have to be adjusted in turns and therefore the complete update of the weights becomes too slow for channels with small coherence time. In contrast, other approaches employing gradient methods update all the weights at the same time. This algorithms include subspace projections and incremental rotations [2], and sign gradient methods [3][4],[5].

II. System model

The system is a discrete time MIMO setup with a single user, N_T antennas at the transmit side and N_R antennas at the receive side. It is also assumed to use FDD and to be frequency flat. Therefore:

$$\mathbf{y} = \mathbf{H}\mathbf{W}\mathbf{x} + \text{noise} \tag{1}$$

where the equation relates the output of transmit side antenna array $\mathbf{x}$ ($N_b \times 1$, which represents a system with N_b beams) undergoing flat frequency fading for one symbol period, to the output of the receive side antenna array $\mathbf{y}$.
The channel matrix $\mathbf{H}$ is an $N_R \times N_T$ complex matrix, and its elements can be correlated or uncorrelated. The transmit power is restricted and normalized to one. The noise is a WSS complex Gaussian process with i.i.d. vector components and the average noise power on each Rx antenna is N_0. The beamforming matrix is an $N_T \times N_b$ complex matrix. For a single beam system with no beamforming applied, $\mathbf{W}$ is $N_T \times 1$ with elements equal to $1/\sqrt{N_T}$.
In general, the feedback message is transmitted once per slot (it contains N_{fb} bits) and each slot has T symbol periods, for which the channel matrix might differ. For WCDMA systems, $N_{\text{fb}} = 1$ and therefore the feedback rate is 1500 Hz.
System model is shown in Figure 1.

III. WCDMA's transmit diversity feedback modes

Two schemes are defined in the standard [1] for $N_T = 2$. Both attempt to maximize the received power, across multipath components and receive antennas. They differ in the quantization of the feedback message and the feedback bit filtering. The system is a single beam system with 2 transmit antennas. Consider the existence of L multipath components. Then the channel coefficients from each transmit antenna and multipath component can be arranged in an $L \times N_T$ matrix $\mathbf{H}_n$, associated to the receive antenna n.
Defining the matrix $\mathbf{R} := \sum_{n=1}^{N_R} \mathbf{H}_n^{\dagger}\mathbf{H}_n$, the problem of choosing the weights $\mathbf{W}$ so that to maximize the received power is:

$$\underset{\mathbf{W}\,:\,||\mathbf{W}||^2=1}{\text{argmax}} \quad \mathbf{W}^{\dagger}\mathbf{R}\mathbf{W} \tag{2}$$

The optimal weight vector is the dominant eigenvector of $\mathbf{R}$. When $N_T = 2$, there is freedom to choose the angle of one

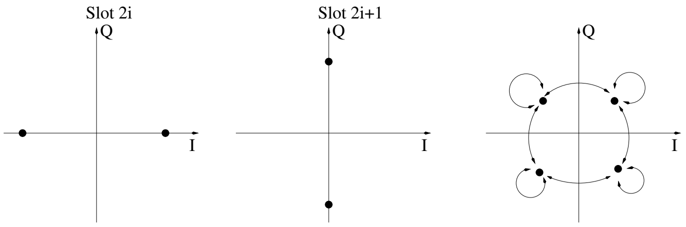

Fig. 2. Weight calculation at BS for Mode 1, with filtering over two slots

of the weights, and for simplicity it is taken as a real number. The UE computes the optimum weights, quantizes them and transmits the resulting bits to the BS through the feedback channel ($FSMph$ field in uplink signal). When the channel exhibits flat fading, the solution is:

$$|w_2| = \frac{|h_2|}{\sqrt{|h_1|^2 + |h_2|^2}} \quad (3)$$

$$\arg(w_2) = \arg(h_2) - \arg(h_1) \quad (4)$$

$$w_1 = \frac{|h_1|}{\sqrt{|h_1|^2 + |h_2|^2}} \quad (5)$$

The beamforming matrix at a given slot is a column vector:

$$\mathbf{W} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \quad (6)$$

Denote b_k the feedback bit coming in slot k.

A. Mode 1

In this mode w_1 is fixed and chosen as $1/\sqrt{2}$. The other weight is updated every time a feedback bit comes. The second weight w_2 is quantized to two bits, therefore it is approximated to elements of a QPSK constellation.

$$w_2[k] = \frac{1}{\sqrt{2}} e^{j\phi[k]} \quad (7)$$

$$\phi[k] = \arg(j^{k \mod 2}\mathrm{sgn}(b_k) + j^{(k-1) \mod 2}\mathrm{sgn}(b_{k-1})) \quad (8)$$

Equation (8) is a "bit filtering" and limits the possible transitions for weight w_2 as a result of the incoming bits. Feedback bits and weight transitions are shown in Figure 2 (taken from [6]).

B. Mode 2

This mode allows for a longer message, where the first bit controls the ratio of the powers being transmitted through the antennas, and the other three control the phase of w_2 (again, w_1 is assumed to be real, but this time the magnitude can change). This means that the quantized weights (for w_2) lie on a constellation consisting of two circles (circular 16QAM), as shown in Figure 3.
The power adjustment, as function of the first feedback bit is done according to Table I. The phase changes are also limited, and given in Table II, where the phase is computed as function of all the current three phase feedback bits.

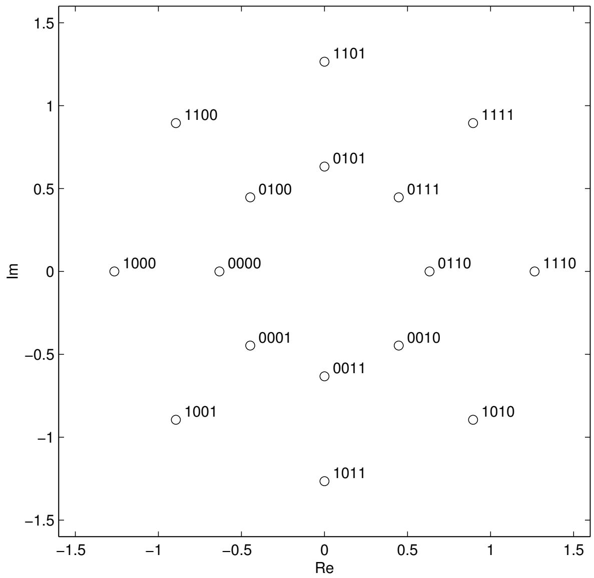

Fig. 3. Weight states in Mode 2, with associated labels (source: [6] and courtesy of authors).

1st bit	Tx1 power	Tx2 Power
0	0.2	0.8
1	0.8	0.2

TABLE I
POWER ADJUSTMENT ACCORDING TO FIRST FEEDBACK BIT IN MODE 2

C. Performance

As expected, the tracking capabilities diminish as the fading rates increase. For Rayleigh fading, analytical results can be derived ([6], section 11.6), taking into account errors in the feedback channel. The increase in the received power due to this adaptive beamforming is shown in Figures 5 and 6.
When the feedback delay is neglected, the BER performance bounds can be computed. Performance will deteriorate as the mobile speed increases and feedback latency is taken into account. Bounds are shown in Figure 4.

IV. SIGNED GRADIENT METHODS

The system under consideration is assumed to use FDD, and then the CSI must come from the closed loop feedback. However, since the feedback rate is limited, the goal of the feedback message is not to make the channel known at the transmitter end, but to assist a distributed search for the optimal beamforming matrix $\mathbf{W}_{\text{opt}}$.
Abstracting the distributed aspect of the system (the BS and UE finding the best weights by sharing information through the closed loop feedback channel) and using $N_b = 1$ (single

Phase (degrees)	180	-135	-90	-45	0	45	90	135
Fb. bits 2,3,4	000	001	011	010	110	111	101	100

TABLE II
PHASE OF SECOND WEIGHT, ACCORDING TO FEEDBACK BITS 2,3,4 IN MODE 2

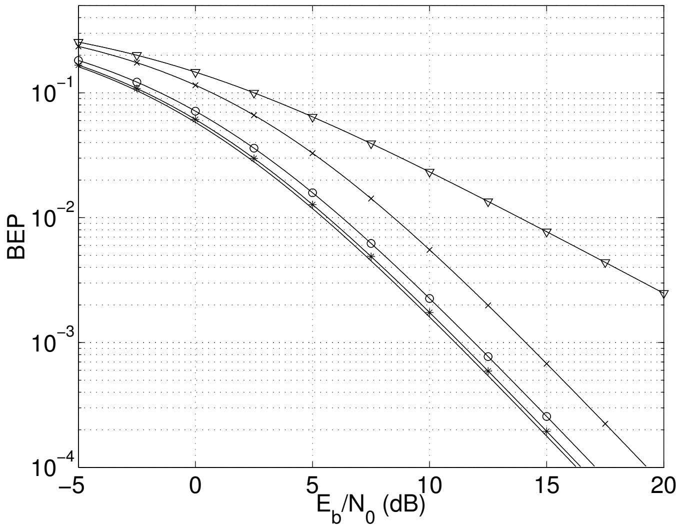

Fig. 4. Performance comparison between Mode 1 (-o-), Mode 2 (-*-) and Space-Time Transmit Diversity (STTD) ($- \times -$) and single antenna transmission ($-\nabla-$). System is 2x1, feedback delay is neglected, channel is block fading for STTD (source: [6] and courtesy of authors).

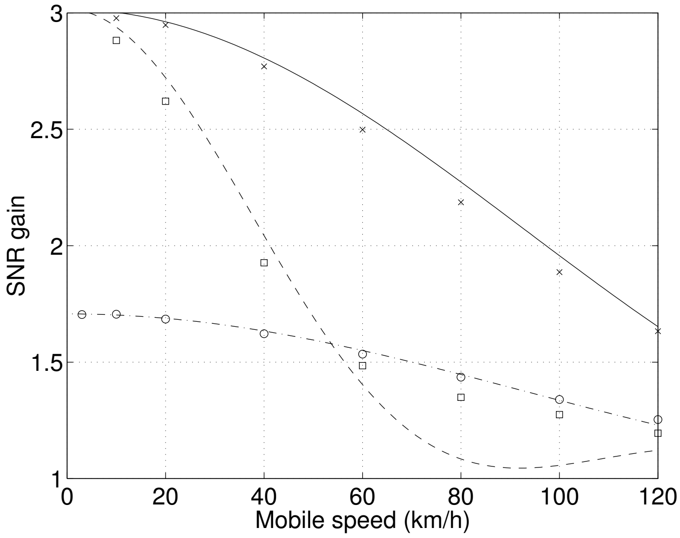

Fig. 5. Received power with adaptive transmit beamforming, no feedback errors. (o) and (-.-) represent Mode 1 (simulation and theoretical result, respectively), cophase algorithm (extension to Mode 1, $N_T = 4$) at 1500 (- -) (simulations: □) and 4500 bps feedback rate (-) (and simulated (x)). Source: [6] and courtesy of authors.

beam system), the basic optimization problem is to maximize $J(\mathbf{W}) := ||\mathbf{HW}||^2$ subject to $||\mathbf{W}||^2 = 1$. Known iterative algorithms perform a gradient based adaptation of the type:

$$\mathbf{W}_{k+1} = \mathbf{W}_k + \beta \mathbf{g}_k \tag{9}$$

where $\mathbf{g}_k$ denotes the current estimate of the gradient of $J(\mathbf{W})$.

Due to limitations in the feedback channel's capacity, it is not possible to feed back the entire gradient vector or the entire instantaneous dominant eigenbeam, because for channels other than slow fading channels, the beam would be outdated and useless by the time it arrived completely at the BS. Therefore,

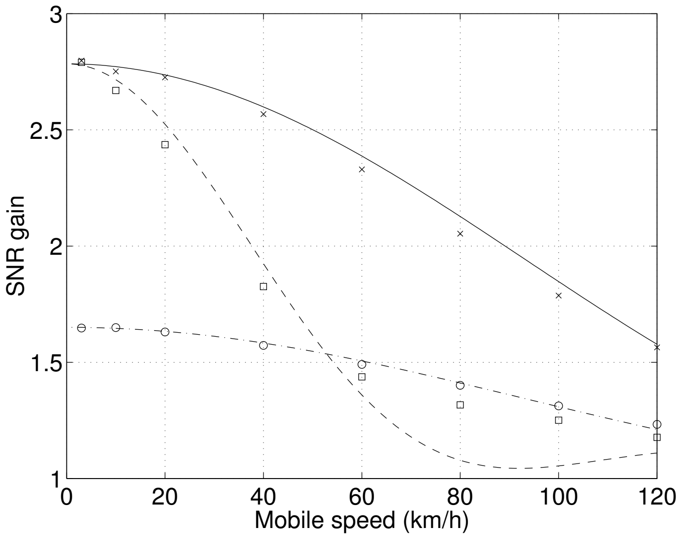

Fig. 6. Received power with adaptive transmit beamforming, 4% feedback errors. (o) and (-.-) represent Mode 1 (simulation and theoretical result, respectively), cophase algorithm (extension to Mode 1, $N_T = 4$) at 1500 (- -) (simulations: □) and 4500 bps feedback rate (-) (and simulated (x)). Source: [6] and courtesy of authors.

the gradient can be quantized and transmitted, or approximated and signalled with reduced rate by exploiting some coordination between transmitter and receiver. In the latter case, when the feedback message length is fixed, the performance of the algorithm will be determined by the quality of the gradient estimate employed in Eq.(9).

Consider update and transmission of the feedback message done once per slot. Therefore, for slot k, the following quantities are defined:

b_k: feedback bit computed in the current slot
$\mathbf{g}_k$: gradient estimate applied in the current slot
$\mathbf{p}_k$: perturbation to be tested in the current slot
$\mathbf{W}_k$: beamforming matrix applied in the current slot
$\hat{\mathbf{R}}_k$: sample estimate of the channel covariance matrix

A general scheme for these signed gradient algorithms ([4], [3],[5]) would be, shown in pseudo-code, at slot k:

1) BS: Update $\mathbf{g}_k$.
2) BS: $\mathbf{W}_k = \mathbf{W}_{k-1} + \beta \mathbf{g}_k$; $\mathbf{W}_k = \mathbf{W}_k / ||\mathbf{W}_k||$
3) BS: Generate new $\mathbf{p}_k$.
4) BS: $\mathbf{W}_e = \mathbf{W} + \beta \mathbf{p}_k$; $\mathbf{W}_0 = \mathbf{W} - \beta \mathbf{p}_k$;
5) UE: $p_e = \mathbf{W}_e^\dagger \hat{\mathbf{R}}_k \mathbf{W}_e / ||\mathbf{W}_e||^2$; $p_o = \mathbf{W}_o^\dagger \hat{\mathbf{R}}_k \mathbf{W}_o / ||\mathbf{W}_o||^2$
6) UE: if ($p_e > p_o$): $b_k = 1$ else: $b_k = -1$
7) UE: transmit b_k, so that it becomes b_{k-1} at next slot
8) BTS: receives b_k
9) Go to first step, b_k becomes b_{k-1}, $\mathbf{g}_k$ becomes $\mathbf{g}_{k-1}$

A. Stochastic gradient approximation

In scheme [3], the perturbation vectors are always random vectors selected i.i.d. by the BS, i.e., $\mathbf{p}_k =$ `new random vector`, drawn every time in an independent fashion. Also, the gradient estimate is only based on the decision fed back by the mobile, i.e. $\mathbf{g}_k = b_{k-1} \mathbf{p}_{k-1}$.

B. Deterministic gradient approximation

In [4], $\mathbf{p}_k$ comes from a fixed set of orthogonal vectors, $\mathbf{p}_k \in \{\mathbf{v}_i\}_{i=1}^{2N_T}$, where the vectors are any orthogonal set and the same set multiplied by the imaginary unit j (e.g., vectors coming from the Discrete Fourier Transform, Discrete Cosine Transform or identity matrices). The gradient update proceeds like in [3].
For example, for $N_T = 2$ and using the DFT vectors:

$$\{\mathbf{v}_i\}_{i=1}^{2N_T} = \left\{ \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \begin{bmatrix} j \\ j \end{bmatrix}, \begin{bmatrix} j \\ -j \end{bmatrix} \right\} \quad (10)$$

C. Filtered gradient

Convergence of signed gradient algorithms can be improved by exploiting the correlation between update directions (the gradient estimates) and by restricting the search directions to be orthogonal to the previous gradient, as in the steepest descent method (this can be done by taking a random vector and subtracting its projection over the previous update direction, like in the Gram Schmidt procedure).
Thus:

- $\mathbf{g}_k = b_{k-1}\mathbf{p}_{k-1} + \lambda_G \mathbf{g}_{k-1}$ (gradient reuse or filtering).
- $\mathbf{p}_k$ = aux - $(\mathbf{g}_{k-1}^{\dagger}\mathtt{aux}/||\mathbf{g}_{k-1}||^2)\mathbf{g}_{k-1}$ where aux is a random vector drawn in an independent fashion. I.e., search direction is orthogonal to the gradient estimate used in the previous slot.

D. Convergence in static channel

As an example, the convergence characteristics of the algorithms are measured by trying to maximize the average received power of a 4x1 channel, which has covariance matrix $\mathbf{R} = \mathbf{R}_{\mathrm{T_x}}^{\mathrm{T}}$, where $\mathbf{R}_{\mathrm{T_x}}$ is the the transmit correlation matrix from 3GPP case 2 (as given in I-Metra scripts described in [7]). This is shown in Figure 7.

E. Bit error probabilities

As an illustration, bit error probabilities (BER) for [3] and [5] are compared, and the benefit of reusing the gradient can be seen. Figure 9 shows a fixed BER as function of speed for a fixed SNR, while Figure 8 shows BER against SNR at selected speeds and parameters.

F. Correlated channels

The correlated channels are generated with a transmit correlation matrix $\mathbf{R}_{\mathrm{T_x}}$, so that:

$$\mathbf{H} = \mathbf{G}\mathbf{B}^{\dagger} \quad ; \quad \mathbf{B}\mathbf{B}^{\dagger} = \mathbf{R}_{\mathrm{T_x}}^{\mathrm{T}} \quad (11)$$

where $\mathbf{G}$ is a 1x4 i.i.d complex Gaussian vector with covariance matrix $\mathbf{I}_4$. The correlation matrix in use is a Hermitian Toeplitz matrix given by:

$$\mathbf{R}_{\mathrm{T_x}} = \begin{bmatrix} 1 & a & b & c \\ a^* & 1 & a & b \\ b^* & a^* & 1 & a \\ c^* & b^* & a^* & 1 \end{bmatrix} \quad (12)$$

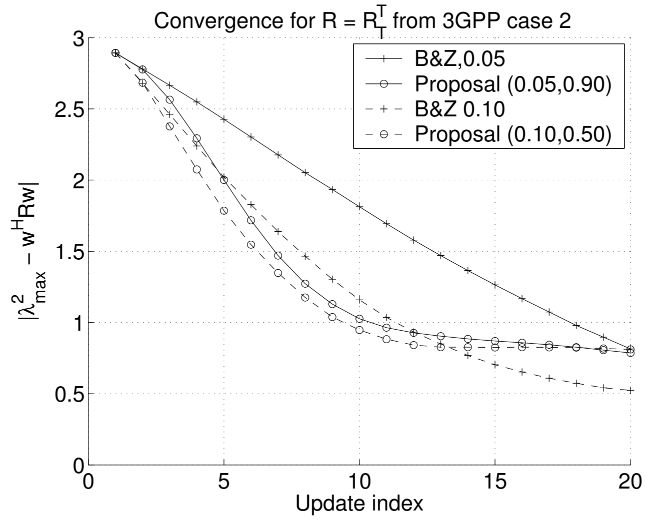

Fig. 7. Convergence in static channel. "B&Z" denotes the algorithm in [3]. Filtered gradient converges uniformly faster when same convergence step (β) is used, up to the point where learning curves cross. Filtered gradient's parameters are given as (β, λ_G). $\lambda_{\max}^2$ is the dominant eigenvalue of $\mathbf{R}$. "Proposal" refers to [5].

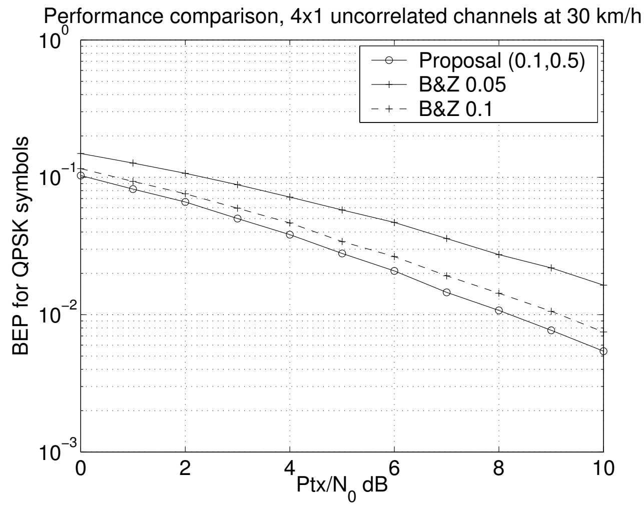

Fig. 8. Bit Error Rate curves for 4x1 uncorrelated channels at 30 km/h. "B&Z" refers to [3], proposal's parameters are given as (β, λ_G) and refer to [5].

For example, the 3GPP case 2 correlation matrix (that corresponds to a strongly correlated scenario), computed through the Matlab scripts described in [7] (spacing between antennas equal to $\lambda/2$) will give $a = 0.4640 + 0.8499j$, $b = -0.4802 + 0.7421j$ and $c = -0.7688 - 0.0625i$. The channel covariance matrix is then $\mathbf{R} = \mathrm{E}\left\{\mathbf{H}^{\dagger}\mathbf{H}\right\} = \mathbf{R}_{\mathrm{T_x}}^{\mathrm{T}}$ and has same eigenvalues as $\mathbf{R}_{\mathrm{T_x}}$, namely 3.7246, 0.2599, 0.0149 and 0.0005. This case from 3GPP corresponds to a macrocell scenario with Laplacian Power Azimuth Spectrum (PAS) and Azimuth Spread (AS) $5°$.
The spatial correlation in the channel moves between two extremes: no correlation, where the elements of $\mathbf{H}$ fade independently, in which case the channel covariance matrix is an identity matrix of size N_T, and full correlation, where all the average power is concentrated in a single eigenvalue

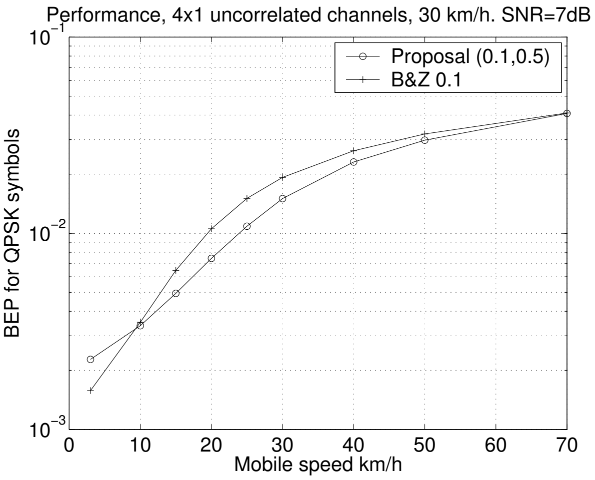

Fig. 9. Bit Error Rate as function of speed in 4x1 uncorrelated channels. "B&Z" refers to [3], proposal's parameters are given as (β, λ_G) and refer to [5]. Transmit power to noise power ratio is 7 dB.

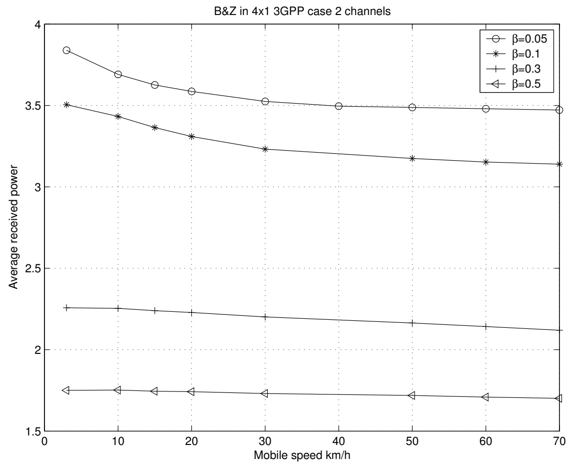

Fig. 10. Average received power as function of mobile speed for algorithm [3]

of the covariance matrix. The first case represents a dominant eigenbeam that is changing in time according to the channel speed. As the correlation increases, this eigenbeam changes less and less in time, meaning that there is more time for the algorithms to converge, and therefore the performance will be dictated by the accuracy of the convergence. In other words, as the correlation increases, there is need for small convergence steps (Figure 10), in order to provide accuracy, and thus the gradient reuse will not help, and λ_G should be set to zero in the fully correlated case.

The amount of correlation can be measured by the eigenvalue spread of the channel covariance matrix: $\rho := 1 - \sigma_{\min}/\sigma_{\max}$ where $\sigma_{\min}$ and $\sigma_{\max}$ represent the extreme eigenvalues of $\mathrm{E}\left\{\mathbf{H}^{\dagger}\mathbf{H}\right\}$. Thus in the uncorrelated case, $\rho = 0$ and $\rho \rightarrow 1$ as correlation increases.

V. Conclusions

An overview of adaptive transmit beamforming based on the closed loop feedback has been presented. Approaches shown here include concepts already present in the current WCDMA standard, as well as proposed techniques aiming to increase the performance in fast fading channels. Emphasis has been made in schemes employing short feedback messages of one bit length, since they have the best chance to improve performance in fast fading channels.

References

[1] 3GPP RAN WG1, "TS25.201: Physical layer - general description," 3GPP, Tech. Rep., June 2000.

[2] W. Utschick and C. Brunner, "Efficient tracking and feedback of DL-eigenbeams in WCDMA," in *Proceedings of the 4th European Mobile Communications Conference*, 2001.

[3] B. C. Banister and J. R. Zeidler, "A simple gradient sign algorithm for transmit antenna weight adaptation with feedback," *IEEE Transactions on Signal Processing*, vol. 51, no. 5, pp. 1156–1171, May 2003.

[4] B. Raghothaman, "Deterministic perturbation gradient approximation for transmission subspace tracking in FDD-CDMA," *IEEE International Conference on Communications, 2003.*, vol. 4, pp. 2450–2454, May 2003.

[5] E. Zacarías B., R. Wichman, and S. Werner, "Filtered gradient algorithm for closed loop mimo systems," *IEEE 61st Semiannual Vehicular Technology Conference (spring)*, May 2005.

[6] A. Hottinen, O. Tirkkonen, and R. Wichman, *Multi-antenna transceiver techniques for 3G and beyond.* John Wiley and Sons, January 2003.

[7] L. Schumacher, "Description of the MATLAB(R) implementation of a MIMO channel model suited for link-level simulations," Available from http://www.ist-imetra.org/, March 2002, version 0.1.