

Research Article

Maximum-Likelihood Semiblind Equalization of Doubly Selective Channels Using the EM Algorithm

Gideon Kutz and Dan Raphaeli

Faculty of Engineering—Systems, Tel-Aviv University, Tel-Aviv 66978, Israel

Correspondence should be addressed to Gideon Kutz, gideon.kutz@freescale.com

Received 5 August 2009; Revised 16 April 2010; Accepted 9 June 2010

Academic Editor: Cihan Tepedelenlioğlu

Copyright © 2010 G. Kutz and D. Raphaeli. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Maximum-likelihood semi-blind joint channel estimation and equalization for doubly selective channels and single-carrier systems is proposed. We model the doubly selective channel as an FIR filter where each filter tap is modeled as a linear combination of basis functions. This channel description is then integrated in an iterative scheme based on the expectation-maximization (EM) principle that converges to the channel description vector estimation. We discuss the selection of the basis functions and compare various functions sets. To alleviate the problem of convergence to a local maximum, we propose an initialization scheme to the EM iterations based on a small number of pilot symbols. We further derive a pilot positioning scheme targeted to reduce the probability of convergence to a local maximum. Our pilot positioning analysis reveals that for high Doppler rates it is better to spread the pilots evenly throughout the data block (and not to group them) even for frequency-selective channels. The resulting equalization algorithm is shown to be superior over previously proposed equalization schemes and to perform in many cases close to the maximum-likelihood equalizer with perfect channel knowledge. Our proposed method is also suitable for coded systems and as a building block for Turbo equalization algorithms.

1. Introduction

Next generation cellular communication systems are required to support high data rate transmissions for highly mobile users. These requirements may lead to doubly selective channels, that is, channels that experience both frequency-selective fading and time-selective fading. The frequency selectivity of the channel stems from the requirement to support higher data rates that necessitates the usage of larger bandwidth. Time selectivity arises because of the need to support users traveling at high velocities as well as the usage of higher carrier frequencies. It is therefore an important challenge to develop high-performance equalization schemes for doubly selective channels.

Doubly selective channels can arise both in single-carrier systems and in Orthogonal Frequency Division Multiplexing (OFDM) systems. In single-carrier systems, the doubly selective channel is modeled as a time-varying filter and introduces time-varying Inter Symbol Interference (ISI). In OFDM systems, the time selectivity of the channel destroys the orthogonality between subcarriers and introduces Inter

Carrier Interference (ICI) while the frequency selectivity of the channel causes the ICI to be frequency varying. In this paper, we concentrate on single-carrier systems only.

The problem of equalization for doubly selective channels has been extensively researched. Several methods for only training-based-equalization were proposed in [1, 2]. Semi-blind equalization methods, that can benefit from both the training and data symbols, were proposed based on linear processing [3, 4] and Decision-Feedback Equalization (DFE) [5]. However, the performance of these equalization methods may not be satisfactory, especially when only one receiving antenna is present [6]. Moreover, the constant advance in processing power calls for more sophisticated equalization schemes that can increase network capacity.

Maximum-likelihood detection based on the Viterbi algorithm is a widely known technique for slowly fading channels. For higher Doppler rates, this method is not satisfactory due to the inherent delay in the Viterbi detector which causes the channel estimator part not to track the channel sufficiently fast. A partial remedy is offered by the Per Survivor Processing (PSP) approach, proposed

originally in [7] and justified theoretically from the Expectation-Maximization (EM) principle in [8]. Using the PSP approach, the channel estimation is updated along each survivor path each symbol period using Least Mean Square (LMS) or Recursive Least Squares (RLS) [9]. However, for high Doppler rates, the performance is limited as these algorithms are not able to track fast fading channels [10]. Improved performance can be gained by using Kalman filtering [9, 11–13] but this approach requires the knowledge of the channel statistics which is normally not known a priori and its estimation will likely not be able to track fast fading. Low-complexity alternatives to the PSP were also proposed [11, 14].

Fast time-varying channel estimation might be achieved using the basis expansion (BE) model [15]. In this method, the channel's time behavior is modeled as a linear combination of basis functions. Basis functions can be polynomials [15], oversampled complex exponentials [2, 16], discrete prolate spheroidal sequences [17], and Karhunen-Loeve decomposition of the fading correlation matrix [10]. Several receiver structures were proposed based on the combination of the BE with Viterbi algorithm variants like PSP [10], M -algorithm [14], and minimum survivor sequence [6, 18]. One common drawback of these methods is that the channel estimation part uses only hard decisions and does not weight the probability of different hypothesis for the symbol sequences. Moreover, all of these methods provide only hard decisions outputs which make them unsuitable for coded systems.

In order to enable the channel estimation part to benefit from soft decisions, several MAP-based algorithms combined with recursive, RLS-based, channel estimation were proposed [19–22]. The combination of MAP decoding and maximum likelihood channel estimation can be justified using the EM principle. This leads to an iterative detection and channel estimation algorithm based on the Baum-Welch (BW) algorithm, proposed in [23] and modified for reduced complexity in [24] for non-time-varying channels (See [25] and references therein for more non-time-varying semiblind equalization methods). Adaptation for doubly selective channels is found in [26] based on incorporation of LMS and RLS in the algorithm.

Iterative MAP detection combined with polynomial BE was proposed in [27]. Unfortunately, this method cannot be directly extended to higher order BE models required in high mobility environments because the choice of polynomial expansion creates numerical difficulties for higher BE models. Furthermore, the equalization and channel estimation in [27] are done in a two-step ad hoc approach which is not a true EM (see Appendix A) and exhibits degraded performance in our simulations.

Finally, Turbo equalization schemes, encompassing iterative detection and decoding, were proposed based on RLS/LMS channel estimation [22, 26] and BE channel estimation [28]. The latter method employs a low-complexity approximation to the MAP algorithm for the detection part. It requires, however, that the channel statistics is fully known a priori.

In this paper, we present a novel method for semi-blind ML-based joint channel estimation and equalization for doubly selective channels. The method is based on an adaptation of the EM-based algorithm for doubly selective channels by incorporating a BE model of the channel in the EM iterations. Using the BE method, we can simultaneously use long blocks thereby enhancing the performance in noisy environments without compromising the ability to track the channel because of the usage of sufficiently high-order BE to model the channel time variations. The proposed method is shown to have superior performance over previously proposed methods with the same block size and number of pilots in the block. Alternatively, it requires a lower number of pilots to achieve the same performance thereby enabling more bandwidth for the information. In addition, it is shown to have good performance for relatively small blocks, which is important if low latency in the communication system is required. The proposed algorithm outputs are the log-likelihood ratios (LLRs) of the transmitted bits, making it ideally suited for coded systems and also suitable as a building block for Turbo equalization algorithm that iterates between detection and decoding stages to improve the performance further. We treat the case of uncorrelated channels paths which is the worst case in terms of number of required BE functions. In Appendix B, we discuss the generalization to correlated paths.

Another contribution of the paper is the determination of a pilot positioning scheme that improves the equalizer's performance. In the context of our proposed algorithm, the main purpose of the pilots is the enablement of sufficient quality initialization of the EM iterations so that the probability of convergence to a local maximum is minimized. To that end, we propose an initialization scheme based on a small number of pilots and find the optimal pilot positioning such that the initial channel parameters guess is as close as possible to the channel parameters obtained assuming perfect knowledge of transmitted symbols (this is the channel estimation expected at the end of the BW iterations). It is shown that the pilot positioning depends on the channel's Doppler. For high Doppler rates, our results indicate that spreading the pilot symbols evenly throughout the block leads to the best initial channel guess. This result is surprising as it is different from previous results where the optimal positioning scheme was found to be spreading of *groups* of pilots whose length depended on the channels delay spread [29, 30]. These previous results, however, were obtained using different criteria and channel model. More importantly, the analysis in these papers was restricted and did not consider pilot groups shorter than the channel's delay spread as done in this paper. Therefore, these previous results do no contradict with our new result.

Pilot positioning was discussed in [31–33] and conditions for MMSE optimality of both the pilot *sequence* and *positioning* were derived. The resulting sequences and positioning are, however, less attractive for practical implementations. This is because most of them require that the pilots and data overlap in time which complicated the receiver structure. The only optimal scheme proposed in [32, 33] with nonoverlapping pilots and data requires a pilot

pattern that results in very high peak to average transmission which is not desirable in practical communications systems. In contrast, we optimize the pilot positioning given a predefined pilot sequence (in this paper we use, as an example, Barker sequence). This allows us to derive optimal pilot positioning for a given pilot sequence that meets some other constraints (e.g., constant envelope signals, low peak-to-average ratio, etc.).

The rest of the paper is organized as follows. In Section 2, we present the system model and introduce the BE model. In Section 3 we present our proposed method for semi-blind joint channel estimation and equalization for doubly selective channels. Section 4 discussed the BE functions set selection. Our results regarding optimal pilot placement are presented in Section 5. Section 6 presents our simulation results and conclusions are drawn in Section 7. Partial results of this work were introduced in a conference paper [34].

2. Problem Formulation

2.1. System Model. The transmitted symbols vector $\mathbf{x} = [x_0, \dots, x_{N-1}]^T$ is an i.i.d sequence with uniform distribution over an arbitrary constellation of size M . The sequence is transmitted over an unknown multipath channel modeled as a time-varying finite impulse response (FIR) filter with coefficients vector at time sample n , $\mathbf{h}_n = [h_{0,n}, \dots, h_{L-1,n}]^T$. The received sample at time n is

$$y_n = \sum_{i=0}^{L-1} h_{i,n} x_{n-i} + w_n = \mathbf{x}_n^T \mathbf{h}_n + w_n, \quad (1)$$

where $\mathbf{y} = [y_0, \dots, y_{N-1}]^T$ is the received vector (observation vector) and $\mathbf{x}_n = [x_n, \dots, x_{n-L+1}]^T$ represents a branch (transition) on the trellis formed by the channel's memory [23]. There are M^L possible branches at each time sample n . Each possible branch is denoted by the row vector $\mathbf{s}_{k,n}$, where $0 \leq k < M^L$ and $0 \leq n < N$. Finally, $\mathbf{w} = [w_0, \dots, w_{N-1}]^T$ is an Additive White Gaussian Noise (AWGN) sequence with zero mean and an unknown variance σ^2 .

The time selectivity of the channel is typically characterized by the normalized Doppler frequency defined as

$$f_{nd} = \frac{T_s f_c v}{c}, \quad (2)$$

where f_c is the communication system carrier frequency, v is the user's velocity, c is the speed of light, and T_s is the time of one symbol.

The sequence $\mathbf{h}^{(i)} = [h_{i,0}, \dots, h_{i,N-1}]^T$, which represents the time variations of the i th channel's path, is modeled as a wide-sense stationary stochastic process with autocorrelation function [35]

$$C_i(\Delta n) = \alpha_i J_0(2\pi f_{nd} \Delta n), \quad (3)$$

where J_0 is the zero-order Bessel function and Δn is the time difference in sample units. Furthermore, α_i is the average power of the i th channel path and the power profile of the channel is $\alpha = [\alpha_0, \dots, \alpha_{L-1}]^T$. In addition, we make the following standard assumptions.

- (A1) Information symbols, channel realization, and noise samples are statistically independent.
- (A2) The channel's paths are statistically independent (uncorrelated scattering [35]).

2.2. Basis Expansion Model. Using the BE approach, we model the time variation of each channel's path with a linear combination of several basis functions, that is, the value of the i th path at time n is

$$h_{i,n} = \sum_q b_n(q) g_{i,q}, \quad (4)$$

where $b_n(q)$ is the n th element from the q th basis and $g_{i,q}$ is the combination coefficient of the i th path and the q th basis function. The advantage of this description is that the complete time and frequency behavior of the channel is described using a relatively small set of LQ coefficients vector $\mathbf{g} = [g_{0,0}, \dots, g_{0,Q-1}, g_{1,0}, \dots, g_{L-1,Q-1}]^T$. We further define a BE matrix as

$$B = [\mathbf{b}_0^T, \dots, \mathbf{b}_{N-1}^T]^T, \quad (5)$$

and $\mathbf{b}_n = [b_n(0), \dots, b_n(Q-1)]$ is a row vector of the function values at time n . Equation (4) in matrix form is then

$$\mathbf{h}^{(i)} = B \mathbf{g}_i, \quad (6)$$

where $\mathbf{g}_i = [g_{i,0}, g_{i,1}, \dots, g_{i,Q-1}]^T$.

3. The Baum-Welch Algorithm for Equalization of Doubly Selective Channels

In this section, we present our new algorithm for semi-blind maximum-likelihood joint channel estimation and equalization for doubly selective channels. We treat channels with uncorrelated paths as this is the worst case in term of number of required basis functions in the BE description. In Appendix B, we extend the algorithm for channels with correlated paths.

3.1. Algorithm for Blind Equalization. If we define the ML estimation of the channel parameters as $\theta = [\hat{\mathbf{g}}^T, \sigma^2]^T$, we would like to find θ such that

$$p(\mathbf{y} | \theta) = \sum_{\mathbf{s} \in S} p(\mathbf{y}, \mathbf{s} | \theta) \quad (7)$$

is maximized. The sum is over all possible transmitted symbols vectors, or equivalently over all possible transition sequences in the trellis S . Direct maximization of $p(\mathbf{y} | \theta)$ is an intractable problem. We can, however, maximize this expression iteratively using the EM algorithm [23]. In each iteration we compute, in the E step, the expectation of the log-likelihood of the complete data conditioned on

the observation and our current estimate of θ . At the l th iteration, this value can be shown to be [23]

$$\begin{aligned} Q(\theta | \theta^{(l)}) &= E[\log p(\mathbf{y}, \mathbf{s} | \theta) | \mathbf{y}, \theta^{(l)}] \\ &= \sum_{\mathbf{s} \in \mathcal{S}} p(\mathbf{s} | \mathbf{y}, \theta^{(l)}) \log p(\mathbf{y}, \mathbf{s} | \theta) \\ &= C + \sum_{n=0}^{N-1} \sum_{k=0}^{M^l-1} p(\mathbf{s}_{k,n} | \mathbf{y}, \theta^{(l)}) \\ &\quad \times \left(-\frac{1}{2} \log(\pi\sigma^2) - \frac{1}{2\sigma^2} |y_n - \mathbf{s}_{k,n} \mathbf{h}_n|^2 \right). \end{aligned} \quad (8)$$

We may express \mathbf{h}_n as

$$\mathbf{h}_n = (\mathbf{I}_L \otimes \mathbf{b}_n) \mathbf{g}, \quad (9)$$

where \mathbf{I}_L is an $L \times L$ identity matrix and the sign “ \otimes ” represents a Kronecker product. In the M step, we find new θ such that this expression is maximized, that is

$$\theta^{(l+1)} = \arg \max_{\theta} Q(\theta | \theta^{(l)}), \quad (10)$$

where l is the iteration index. We may now use the new time-varying expression for the channel to get the doubly selective version of the algorithm in [23]. Plugging (9) in (8), repeating the derivation in [23], and utilizing the Kronecker product properties, the resulting update equations are

$$\begin{aligned} \hat{\mathbf{g}}^{(l+1)} &= \left(\sum_{n=0}^{N-1} \left(\sum_{k=0}^{M^l-1} p(\mathbf{s}_{k,n} | \mathbf{y}, \theta^{(l)}) (\mathbf{s}_{k,n}^H \mathbf{s}_{k,n}) \right) \otimes (\mathbf{b}_n^H \mathbf{b}_n) \right)^{-1} \\ &\quad \times \left(\sum_{n=0}^{N-1} \left(\sum_{k=0}^{M^l-1} p(\mathbf{s}_{k,n} | \mathbf{y}, \theta^{(l)}) y_n^* \mathbf{s}_{k,n} \right) (\mathbf{I}_L \otimes \mathbf{b}_n) \right)^H, \end{aligned} \quad (11)$$

where we have used the identity $x\mathbf{y} \otimes z\mathbf{w} = (x \otimes z)(\mathbf{y} \otimes \mathbf{w})$ and

$$(\sigma^2)^{(l+1)} = \sum_{n=0}^{N-1} \sum_{k=0}^{M^l-1} p(\mathbf{s}_{k,n} | \mathbf{y}, \theta^{(l)}) |y_n - \mathbf{s}_{k,n} (\mathbf{I}_L \otimes \mathbf{b}_n) \mathbf{g}|^2. \quad (12)$$

The values of $p(\mathbf{s}_{k,n} | \mathbf{y}, \theta^{(l)})$ are efficiently computed using the Bahl-Cocke-Jelinek-Raviv (BCJR) algorithm [36]. For non-time-varying channels, we have $\mathbf{b}_n = 1$ and (11) reduces to equation (11) in [23]. Our algorithm may also be extended to channels with correlated paths. In this case, the number of BE parameters used for describing all channel's paths may be reduced. In that sense, uncorrelated channels paths may be considered as the worst case. The correlated case is discussed in Appendix B.

3.2. Adaptation to the Semiblind Case. Adaptation of the above algorithm to the semi-blind case where we have some known pilot symbols is straightforward. The only required change is in the computation of $p(\mathbf{s}_{k,n} | \mathbf{y}, \theta^{(l)})$ using the

BCJR algorithm. We modify the branch metrics so that all transitions that are not consistent with known pilots are assigned zero probability. This ensures that transition probabilities are calculated with the a priori information about the pilots.

3.3. Initialization of the Algorithm. Optimization of the EM objective function (8) is a nonlinear process that may converge to a local maximum. It is therefore important to calculate good initial guess for the channel parameters so that the probability of convergence to a local maximum is minimized. We suggest using the available pilot symbols for finding initial channel parameters using the following method. First, we run the BCJR algorithm where the branch metrics are initialized without any initial channel guess by assigning zero a priori probability to all transitions that are not consistent with the known pilots and equal (nonzero) a priori probability to all transitions that are consistent with the pilots. This initialization of the branch metrics represents our best a priori knowledge about the transitions probability in the trellis. From the BCJR algorithm we find $p(\mathbf{s}_{k,n})$ and then use them in (11) to obtain an initial guess for the channel BE parameters. More details on this initialization method can be found in Appendix C.

An important feature of this initialization scheme is that all observations that have some content of pilots in them are taken into account including those with mixed pilot and data contributions. This is in contrast to most other pilot based estimations that take into account observations based on pilots only [30]. This fact turns out to be significant when we discuss how to position the pilot symbols in the block in Section 5.

It should be noted that the above algorithm requires initial synchronization stage to ensure that all major channels taps fall within the searched multipath window. This synchronization stage can be done at a much lower rate than channel estimation update, as the channel tap positions typically drift at a much slower rate compared to the fading rate, and therefore its complexity is negligible. The synchronizations stage is outside the scope of this paper and we assume perfect synchronization throughout the paper.

3.4. Computational Complexity. The computational complexity of the updating equations (11) and (12) is analyzed in Table 1, where we have broken the calculation to several stages and counted the number of complex Multiply-And-Add operations (MAC) for each stage. For comparison, the equivalent complexity of [23] can be obtained from the same table by eliminating the stages for calculating T_2 , T_4 , and \mathbf{h}_n . For (11), it can be seen that for the typical case of $M^l > Q^2/2$, the stages of calculating $T_{1,n}$ and $T_{3,n}$ are more computationally complex than the stages of calculating T_2 and T_4 , respectively. For (12), it can be seen that the second stage of calculating σ^2 is more complex than the first stage of calculating \mathbf{h}_n . Finally, note that all the computationally complex stages of $T_{1,n}$, $T_{3,n}$, and σ^2 do not depend on Q and therefore their complexity is the same as in [23]. We may therefore conclude that the proposed algorithm extends the

TABLE 1: Computational complexity summary.

Equation	Stage	MAC operations
Update channel estimate	$T_{1,n} = \sum_{k=0}^{M^L-1} p(\mathbf{s}_{k,n} \mathbf{y}, \theta^{(l)}) \mathbf{s}_{k,n}^H \mathbf{s}_{k,n}$	$NM^L(L^2/2 + L/2)$
	(11) $T_2 = \sum_{n=0}^{N-1} T_{1,n} \otimes (\mathbf{b}_n^H \mathbf{b}_n),$ $T_{3,n} = \sum_{k=0}^{M^L-1} p(\mathbf{s}_{k,n} \mathbf{y}, \theta^{(l)}) \mathbf{y}_n^* \mathbf{s}_{k,n},$ $T_4 = (\sum_{n=0}^{N-1} T_{3,n} (\mathbf{I}_L \otimes \mathbf{b}_n))^H$ $T_2^{-1} T_4,$	$N(L^2 Q^2/2 + LQ/2)$ $2NM^L L$ NLQ $O(L^3 Q^3)$
Update noise variance	$\mathbf{h}_n = (\mathbf{I}_L \otimes \mathbf{b}_n) \mathbf{g}$	NLQ
(12)	$\sum_{n=0}^{N-1} \sum_{k=0}^{M^L-1} p(\mathbf{s}_{k,n} \mathbf{y}, \theta^{(l)}) \mathbf{y}_n - \mathbf{s}_{k,n} \mathbf{h}_n ^2$	$NM^L(L+2)$

Baum-Welch algorithm [23] for doubly selective channels with only minor increase in complexity.

4. Selection of the Basis Functions

Although many basis functions are possible, previous papers concentrated mostly on three types of basis function sets. The first one is the complex exponentials functions set [2, 37]. The value of the q th basis function of this set at time n is

$$b_n(q) = \exp\left(\frac{j2\pi qn}{N_{\text{bem}}}\right). \quad (13)$$

These functions are periodic with period N_{bem} . In order to avoid modeling errors at the block edges, we therefore set $N_{\text{bem}} = 2N$. The second type of basis functions is [15]

$$b_n(q) = n^q. \quad (14)$$

The functions in (14) model the channel time behavior as polynomial in time. This choice of basis functions may be regarded as a generalization of the channel description in [27] where it was suggested to use first- and second-order polynomials to model the channel time variations.

The best basis functions are the ones that minimizes the mean square error of the fading process description given a finite set of Q basis functions. That is,

$$B = \arg \min_B E |\mathbf{h} - B\mathbf{g}|^2 \quad \text{s.t.} \quad \text{rank}[B] = Q, \quad (15)$$

where the vector \mathbf{h} represents the channel time variation. The solution for this problem is readily available by usage of the Karhunen-Loeve Transform (KLT) [10] and the basis functions are the eigenvectors of the autocorrelation matrix of the Rayleigh fading process. The element n_1, n_2 in the autocorrelation matrix is

$$[R_{\text{corr}}]_{n_1, n_2} = C_i(|n_1 - n_2|). \quad (16)$$

Out of all eigenvectors, the Q vectors that correspond to the largest eigenvalues are selected as the basis set. The

target function in (15) is suitable for flat fading channel. For frequency-selective channel, the mean square error will be simply the sum of the mean square errors of the individual paths and therefore the same solution is optimal for multipath channels. We note that a similar argument is given in [38].

An obvious alternative to the equalization approach we propose in this paper is to divide the data block into small subblocks such that the channel can be considered approximately constant within a subblock period and then equalize each subblock separately using the Baum-Welch algorithm for non-time-varying channels [23]. Interestingly, this subblock scheme can be considered as an instance of the BE approach if we choose Q basis functions for Q subblocks where the q th basis function is equal to one in the symbols time that correspond to the q th subblock and zero elsewhere, that is $B = I_Q \otimes \mathbf{1}_{N/Q}$, where $\mathbf{1}_x$ is a vector on x ones. To justify our approach, we would like to compare it to this subblock approach.

The efficiency of a given set of basis functions may be evaluated by calculating the mean square error of the fading process representation using this set of functions:

$$\begin{aligned} E |\mathbf{h} - B\mathbf{g}|^2 &= E \left[\mathbf{h}^H \left(I - B(B^H B)^{-1} B^H \right) \mathbf{h} \right] \\ &= \text{Tr} \left(R_{\text{corr}} \left(I - B(B^H B)^{-1} B^H \right) \right), \end{aligned} \quad (17)$$

where E and Tr are expectation and matrix trace operators, respectively. Figure 1 plots the required number of basis functions (rank of B) so that the mean square error in (17) is lower than 1% error. As expected, using the eigenvectors as basis functions leads to the lowest number of functions. The polynomial basis set is shown to be quite close to the optimal eigenvectors solution for low normalized Doppler while for high Doppler rates, it is more beneficial to use the complex exponentials basis. The sub-block-based basis functions performance is much worse. This is not surprising as these basis functions do not utilize the correlation between subblocks and force a noncontinuous description of the channel in contrast to the channel's typical behavior. The results shown in Figure 1 confirm that this choice of basis functions is not suitable for Rayleigh fading and provides an explanation to the degraded performance of the subblock method shown in the simulation results section.

5. Placement of Pilot Symbols

5.1. Pilot Positioning Problem Formulation. Pilot placement may influence the equalization performance significantly. Traditionally, pilots have been grouped in big clusters. Recent results, however, indicate that using small groups of pilots that are spread evenly throughout the data block is a better strategy [29, 30, 39]. Proper pilot placement for EM based algorithms is particularly important because of the highly nonlinear nature of the EM objective function in doubly selective channels, which results in many local maxima. The purpose of the pilots is, therefore, to enable sufficient quality channel parameters vector initialization so that the

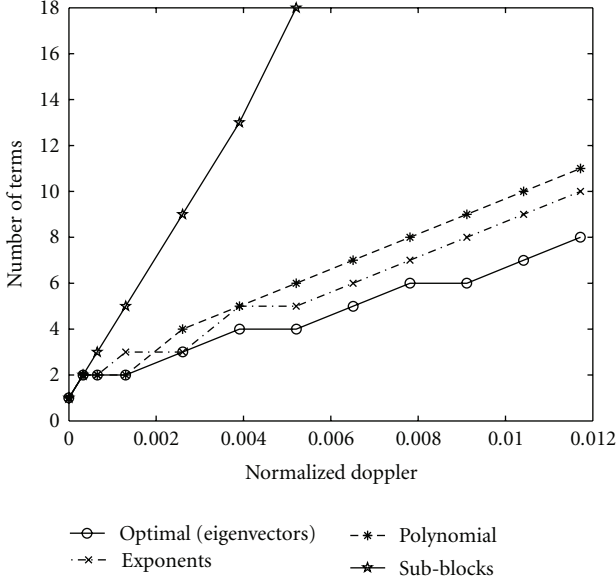


FIGURE 1: Required number of basis functions for mean square error less than 1%. Block size = 256.

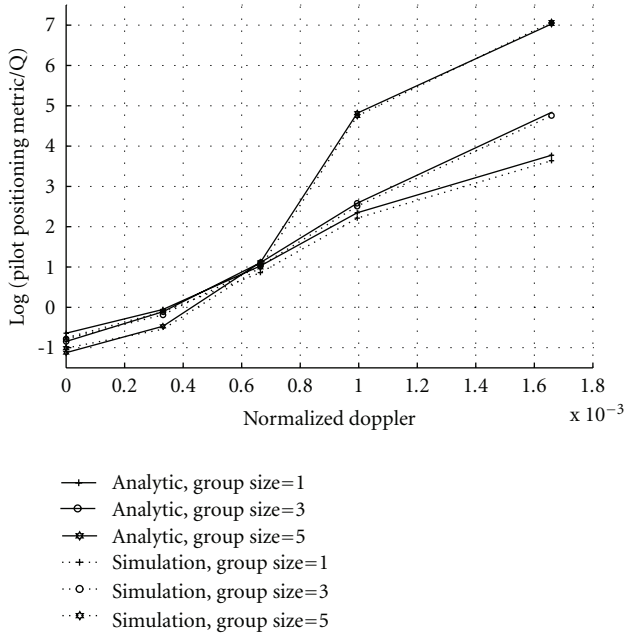


FIGURE 2: Pilot positioning metric (the value to be minimized in (37)) for various L , block size = 512, 5% pilots in the block, channel order $L = 3$, equal average energy paths, number of basis functions $Q = 2\lceil N_{\text{bem}} f_{nd} \rceil + 1$.

probability of convergence to a local maximum is minimized. First, we reformulate the initialization scheme in Section 3.3 as an equivalent Least Squares (LS) problem. Consider first the case where all transmitted symbols are known. In this case, the channel parameters \mathbf{g} can be found with an LS solution to the problem

$$\mathbf{y} = \mathbf{A}\mathbf{g} + \mathbf{w}, \quad (18)$$

where \mathbf{A} represents the known transmitted symbols and the BE model-based time variations. More specifically,

$$\mathbf{A} = \mathbf{X}\mathbf{B}_L, \quad (19)$$

where $\mathbf{X} = [\mathbf{X}_0, \dots, \mathbf{X}_{L-1}]$ and \mathbf{X}_n is an $N \times N$ diagonal matrix such that $\mathbf{X}_n = \text{diag}(x_{-n}, x_{-n+1}, \dots, x_{N-1-n})$ and negative indexes represent data symbols from the previous block that are affecting the observations of the current block due to ISI (if no interblock interference is assumed, these can be replaced by zeros). In addition, $\mathbf{B}_L = \mathbf{I}_{L \times L} \otimes \mathbf{B}$ and $\mathbf{I}_{L \times L}$ is an $L \times L$ identity matrix. The LS solution to (18) is

$$\mathbf{g} = (\mathbf{A}^H \mathbf{A})^{-1} \mathbf{A}^H \mathbf{y}. \quad (20)$$

Now consider the case where only part of the transmitted symbols are known (pilots) and replace the unknown symbols in \mathbf{X} with zeros. The solution to this ‘‘sparse LS’’ problem is

$$\mathbf{g}_p = (\mathbf{A}_p^H \mathbf{A}_p)^{-1} \mathbf{A}_p^H \mathbf{y}_p, \quad (21)$$

where

$$\mathbf{A}_p = \mathbf{X}_p (\mathbf{I}_{L \times L} \otimes \mathbf{B}) = \mathbf{X}_p \mathbf{B}_p \quad (22)$$

and \mathbf{X}_p is defined similarly to \mathbf{X} with nonpilot symbols set to zero. Finally, \mathbf{B}_p is received by setting to zero all elements in the rows corresponding to nonpilot symbols in the matrix \mathbf{B}_L . In Appendix C, we show that the initialization method in Section 3.3 is equivalent to (21). The initialization method is thus equivalent to finding the best BE model parameters vector that fits, in the LS sense, the transmitted pilot sequence (Note that the noise term in this model is not white (since the data is treated as part of the noise). Therefore, a better initialization would be to use weighted least squares method. To do that, however, the noise level and average channel profile need to be known or estimated). Our goal is to position the pilots such that the initial channel guess, based on these pilots, will be optimal according to some criterion. Two reasonable criteria for pilot positioning are

$$\mathbf{p} = \arg \min_{\mathbf{p}} \left[\max_{\mathbf{h}, \mathbf{x}, \mathbf{w}} \|\mathbf{y} - \mathbf{A}\mathbf{g}_p\|^2 \right], \quad (23)$$

$$\mathbf{p} = \arg \min_{\mathbf{p}} E \left[\|\mathbf{g} - \mathbf{g}_p\|^2 \right], \quad (24)$$

where \mathbf{p} is a vector of the pilot positions in the block. The maximum function in (23) and expectation in (24) are taken with respect to the data symbols, noise, and channel realizations. Using these criteria, it might be possible to optimize both the pilots *positions* and the pilot *patterns*. We, however, select known pilot patterns (e.g., Barker sequences) so that we keep constant envelope signals and optimize the positioning for this given pilot pattern. The usage of these two criteria is detailed in the next sections. Interestingly, both criteria lead to the same positioning scheme for high Doppler rates.

5.2. *Worst Case Analysis for Flat Fading Channels.* In this section, we find the best positioning scheme by using (23). First, notice that the criterion may be decomposed to two terms because

$$\begin{aligned} \|\mathbf{y} - \mathbf{A}\mathbf{g}_p\|^2 &= \|\mathbf{y} - \mathbf{A}\mathbf{g} + \mathbf{A}\mathbf{g} - \mathbf{A}\mathbf{g}_p\|^2 \\ &= \|\mathbf{y} - \mathbf{A}\mathbf{g}\|^2 + \|\mathbf{A}\mathbf{g} - \mathbf{A}\mathbf{g}_p\|^2, \end{aligned} \quad (25)$$

where the second equality is justified because, by construction of \mathbf{g} , the term $\mathbf{y} - \mathbf{A}\mathbf{g}$ is orthogonal to the span of the matrix \mathbf{A} to which $\mathbf{A}\mathbf{g} - \mathbf{A}\mathbf{g}_p$ belongs. Note that only the second term is dependent on the pilot positions. Obviously, the best pilot positioning is dependent on the channel and noise realizations. Our goal is to obtain positioning scheme suitable for all channels, data, and noise realizations by optimizing the positioning scheme with respect to the worst case realizations. Using (19) and (22), the second term in (25) may be bounded by

$$\|\mathbf{A}\mathbf{g} - \mathbf{A}\mathbf{g}_p\|^2 = \|\mathcal{A}_p \mathbf{y}\|^2 \leq \sigma_{\max}^2 \|\mathbf{y}\|^2, \quad (26)$$

where $\mathcal{A}_p = A(A^H A)^{-1} A^H - A(A_p^H A_p)^{-1} A_p^H$ and σ_{\max}^2 is the largest eigenvalue of the matrix $\mathcal{A}_p^H \mathcal{A}_p$. For flat fading channels (and any PSK constellation) $X^H X = I$ and, therefore,

$$\begin{aligned} \mathcal{A}_p &= X \left(B(B^H B)^{-1} B^H - B(B_p^H B_p)^{-1} B_p^H \right) X^H \\ &\equiv X \mathcal{B}_p X^H, \end{aligned}$$

$$\text{eig}[\mathcal{A}_p^H \mathcal{A}_p] = \text{eig}[X \mathcal{B}_p^H X^H X \mathcal{B}_p X^H] = \text{eig}[\mathcal{B}_p^H \mathcal{B}_p], \quad (27)$$

where $\text{eig}[D]$ is the vector of eigenvalues of the matrix D . The second equality follows from the fact that for flat fading channels $X^H = X^{-1}$ and $\text{eig}[X^{-1} D X] = \text{eig}[D]$ (Assume that β is eigenvalue of D , that is, $D\mathbf{u} = \beta\mathbf{u}$, define $\mathbf{v} = X^{-1}\mathbf{u}$, then $D X \mathbf{v} = X \beta \mathbf{v}$ and $(X^{-1} D X) \mathbf{v} = \beta \mathbf{v}$. Matrices D and $X^{-1} D X$ have therefore the same eigenvalues).

It follows that minimization of the worst case MSE is achieved by finding a pilot positions vector \mathbf{p} such that

$$\mathbf{p} = \arg \min_{\mathbf{p}} \max \sigma_{\max}^2 = \arg \min_{\mathbf{p}} \max \left(\text{eig}[\mathcal{B}_p^H \mathcal{B}_p] \right). \quad (28)$$

The matrix $\mathcal{B}_p^H \mathcal{B}_p$ is a deterministic function of the BE functions, block size, pilot positioning, and the pilot pattern (sequence). It is therefore possible to find the best positioning scheme for the desired block size, BE model, and pilot sequence with a computer search. For simplicity, we limit the search for patterns in which the pilots are grouped in groups of length \bar{L} and these groups are spread throughout the block as evenly as possible. This means that the pilot positioning we find with this limited search is only optimal amongst all positioning with evenly spaced pilot clusters. However, all previous works on pilot positioning arrived at positioning schemes that are consistent with this structure. It turns out that the best positioning scheme is obtained with

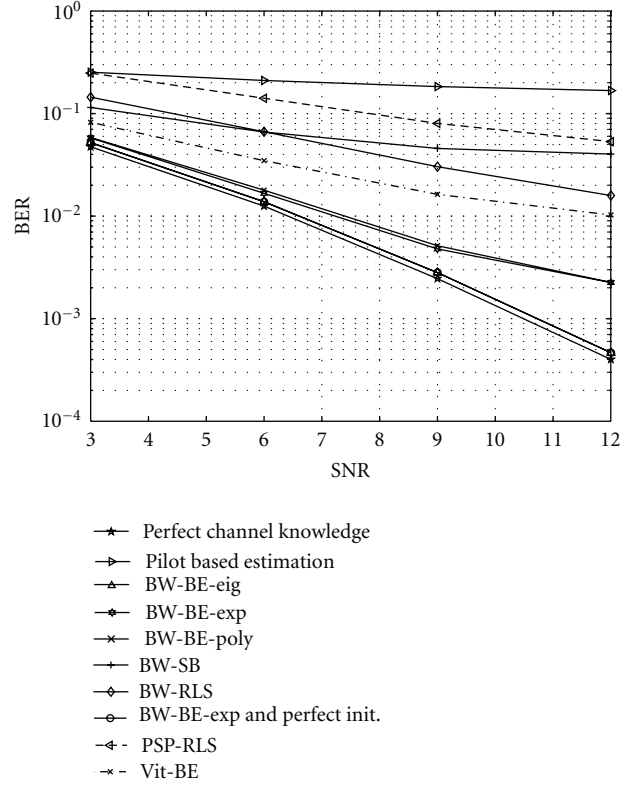


FIGURE 3: Performance of various equalization schemes. Block size = 256, number of pilots = 20, pilot positioning scheme: $\bar{L} = 1$, channel profile = [0 -3 -3] dB.

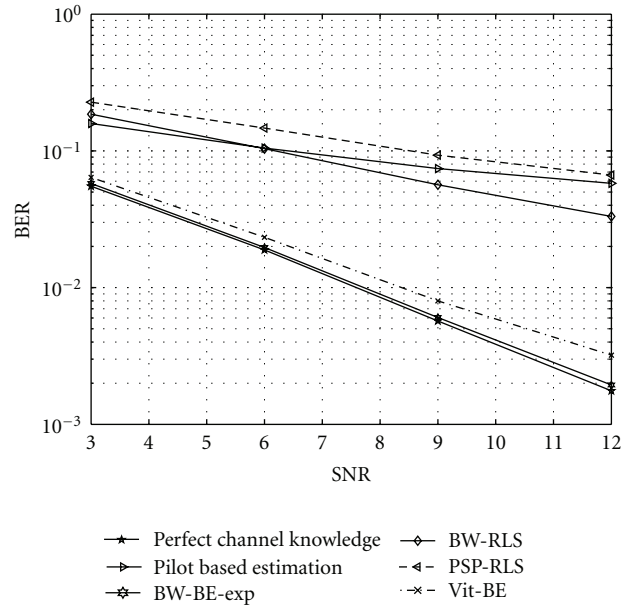


FIGURE 4: Performance of various equalization schemes. Block size = 256, number of pilots = 20, pilot positioning scheme: $\bar{L} = 1$, channel profile = [0 0] dB.

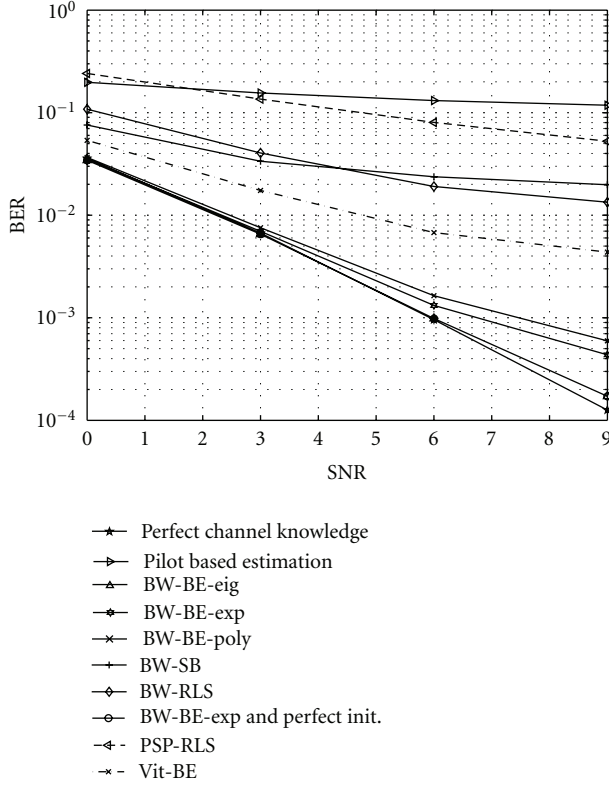


FIGURE 5: Performance of various equalization schemes. Block size = 256, number of pilots = 20, pilot positioning scheme: $\bar{L} = 1$, channel profile = $[0 \ 0 \ 0 \ 0]$ dB.

$\bar{L} = 1$ for all tested block sizes. It is interesting to note that this result is identical to the result in [29] which was obtained using different channel model and criterion.

5.3. Mean Case Analysis for Frequency Selective and Frequency Flat Channels. In this section, we optimize (24). We begin with the approximation

$$B_L^H X^H X B_L \approx C_x, \quad (29)$$

where

$$[C_x]_{kQ+q_1, jQ+q_2} \equiv \begin{cases} \left[\sum_{n=k}^{N-1} \mathbf{b}_n^H \mathbf{b}_n \right]_{q_1, q_2}, & k = j, \\ \left[\sum_{n=0}^{N-1} \mathbf{b}_n^H \mathbf{b}_n x_p^*(n-k) x_p(n-j) \right]_{q_1, q_2}, & k \neq j, \end{cases} \quad (30)$$

and $x_p(m)$ is defined as

$$x_p(m) = \begin{cases} x_m, & m \in \mathcal{P}, \\ 0, & \text{otherwise.} \end{cases} \quad (31)$$

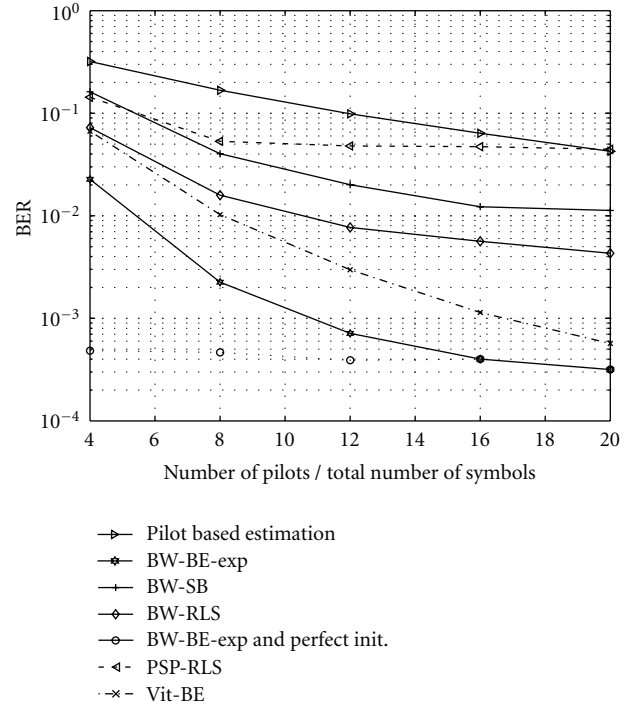


FIGURE 6: Performance of various equalization schemes as a function of the pilot percentage in the block. Block size = 256, SNR = 12 dB, pilot positioning scheme: $\bar{L} = 1$, channel profile = $[0 \ -3 \ -3]$ dB.

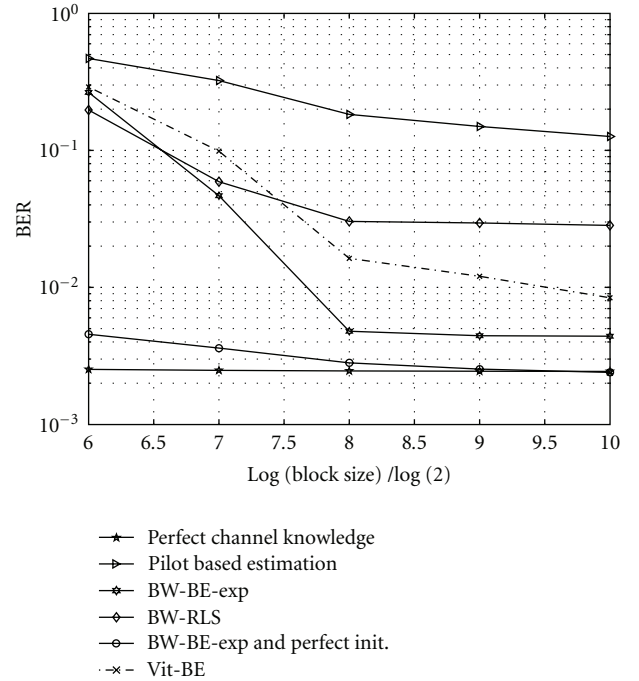


FIGURE 7: Performance of various equalization schemes as a function of the block size. Pilot percentage = 8%, SNR = 9 dB, pilot positioning scheme: $\bar{L} = 1$, channel profile = $[0 \ -3 \ -3]$ dB.

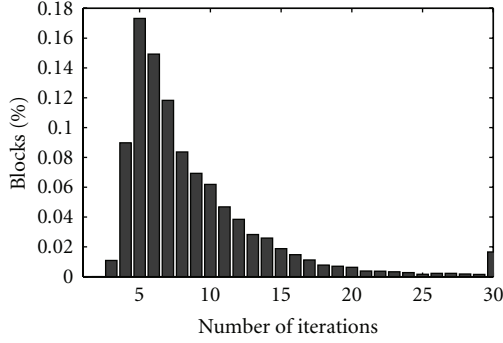


FIGURE 8: Number of iterations required for convergence with block size = 256, SNR = 12 dB, channel profile = [0 -3 -3] dB.

Note that an accurate expression (with no approximation) may be obtained by replacing $x_p^*(n-k)x_p(n-j)$ with $x_{n-k}^*x_{n-j}$. When either x_{n-k} or x_{n-j} is an information symbol (not a pilot), this multiplication result is a random variable, uniformly distributed over a finite set of values with zero average. As a result, for long enough blocks, the contributions from the information symbols to the sum in (30) cancel out and this approximation is fairly accurate. Our criterion may be therefore approximated with

$$\begin{aligned} & \left| \mathbf{g} - \mathbf{g}_p \right|^2 \\ & \approx \left| \left(C_x^{-1} B_L^H - (B_L^H X_p^H X_p B_L)^{-1} B_p^H \right) X^H \mathbf{y} \right|^2 \quad (32) \\ & \equiv \left| \mathcal{D}_p X^H \mathbf{y} \right|^2. \end{aligned}$$

The analysis that follows should be considered valid only for large enough block sizes where (29) is accurate. The expectation of the approximated metric is

$$\begin{aligned} E \left[\left| \mathbf{g} - \mathbf{g}_p \right|^2 \right] &= E \left[\left| \mathcal{D}_p X^H \mathbf{y} \right|^2 \right] \\ &= E \left[\mathbf{y}^H X \mathcal{D}_p^H \mathcal{D}_p X^H \mathbf{y} \right] \quad (33) \\ &= E \left[\text{Tr} \left(X^H \mathbf{y} \mathbf{y}^H X \mathcal{D}_p^H \mathcal{D}_p \right) \right] \\ &= \text{Tr} \left(E \left[X^H \mathbf{y} \mathbf{y}^H X \right] \mathcal{D}_p^H \mathcal{D}_p \right). \end{aligned}$$

The autocorrelation matrix $R = E[X^H \mathbf{y} \mathbf{y}^H X]$ is composed of $L \times L$ submatrices, where the k, j submatrix is $X_k^H \mathbf{y} \mathbf{y}^H X_j$. Using the standard assumption that the channel's paths are statistically independent (assumption A2), we may express the autocorrelation matrix R as a linear combination of the contributions of the channel paths, that is,

$$R \equiv E[X^H \mathbf{y} \mathbf{y}^H X] = \sum_{i=0}^{L-1} R_i + \sigma^2 I_{NL} \quad (34)$$

Using assumptions A1-A2 and (3), the entry $n1, n2$ in the submatrix k, j (or equivalently, the element $kN+n1, jN+n2$ in the matrix R_i) is

$$\begin{aligned} & [R_i]_{kN+n1, jN+n2} \\ &= E \left[h_{i,n1} h_{i,n2}^* \right] E \left[x_{n1-i} x_{n2-i}^* x_{n1-k}^* x_{n2-j} \right]. \quad (35) \end{aligned}$$

where

$$\begin{aligned} E \left[h_{i,n1} h_{i,n2}^* \right] &= \alpha_i J_0 \left(\frac{2\pi f_c v |n1 - n2| T_s}{c} \right), \\ E \left[x_{n1-i} x_{n2-i}^* x_{n1-k}^* x_{n2-j} \right] &= \begin{cases} 1, & n1 = n2, k = j, \\ 1, & k = j = i, n1 \neq n2, \\ x_p^*(n1-k) x_p(n2-j), & n1 = n2, k \neq j, \\ x_p^*(n2-i) x_p(n2-j), & i = k \neq j, \\ x_p(n1-i) x_p^*(n2-i), & n1-k = n2-j, n1 \neq n2, \\ x_p(n1-i) x_p^*(n1-k), & i = j \neq k, \\ x_p(n1-i) x_p^*(n2-i) x_p^*(n1-k) x_p(n2-j), & \text{otherwise.} \end{cases} \quad (36) \end{aligned}$$

The best pilot positioning scheme is therefore

$$\mathbf{p} = \arg \min_{\mathbf{p}} \text{Tr} \left(\left(\sum_{i=0}^{L-1} R_i + \sigma^2 I \right) \mathcal{D}_p^H \mathcal{D}_p \right). \quad (37)$$

This expression is deterministic and depends only on the BE functions, block size, noise variance, channel order (L),

Doppler rate, and pilot sequence. It is therefore possible to find the best pilot positioning for a particular set of parameters by evaluating (37) for various \mathbf{p} . As we did in the previous section, we limit the positioning patterns for patterns in which the pilots are grouped in groups of length \bar{L} , and these groups are spread evenly throughout

the block. This positioning strategy coincides with the pilot positioning in [39] for $\bar{L} = 1$, with the pilot positioning in [30] for $\bar{L} = L$ and with the pilot positioning in [29] for $\bar{L} = 2L + 1$. In addition, every group of pilots is a Barker sequence of length \bar{L} . Barker sequences are known to enable good channel estimation because of their autocorrelation properties. Define the positioning metric as the value to be minimized in (37). A typical behavior of this positioning metric is shown in Figure 2 (based on (37) and in agreement with simulation results).

The optimal positioning strategy is shown to be dependent on the Doppler rate and the number of pilots in the block. As can be seen from Figure 2, for low Doppler rates it is better to use group of pilots as also indicated by [29, 30] (although the difference is not very significant, at least for short delay spreads). For high Doppler rates and a small number of pilots, however, it turns out that using $\bar{L} = 1$ leads to much better results. This is because there is a tradeoff between accurate estimation of the multipath at specific points in time (that is better achieved by grouping the pilots) and tracking the channel time variations (that is better achieved by spreading the pilots throughout the block). Our results indicate that for high velocities using $\bar{L} = 1$ leads to a lower metric value as this means better ability to track time variations. Note that this result is obtained for severe ISI channel with three equal energy paths (and similar result was obtained for channel with 5 equal energy paths). We have also simulated channels with less severe ISI (that is, decaying power profiles), and the advantage of using $\bar{L} = 1$ was even larger, as could be expected. The switching point (Doppler rate beyond which it is advantageous to use $\bar{L} = 1$) is dependent mainly on the percentage of pilots in the block. For larger number of pilots, the switching point will occur at higher Doppler rate. The reason is that for large number of pilots there will be sufficient number of groups in the block to allow tracking of path time variations even when the group size is kept $2L + 1$, so both multipath profile and time variations could be estimated accurately. We, however, are interested in the smallest number of pilots that enables good performance, and in these conditions, $\bar{L} = 1$ is advantageous even for moderate Doppler rates (see Figure 2). This conclusion is somewhat surprising as it is different from previous conclusions in [29, 30]. However, these previous works used different channel models and performance criteria. Moreover, both works considered only pilot groups equal to $2L + 1$ [29] or L [30] or longer, to facilitate their analysis.

6. Simulation Results

6.1. Performance of the Proposed Equalization Scheme. Next, we present simulation results for our proposed equalization scheme. We use a sequence of 2^{17} QPSK symbols that is sent through a doubly selective channel as described in Section 2.1. The normalized Doppler frequency is $f_{nd} = 0.002$, and coherence time, defined as the time over which the channels response to a sinusoid, has a correlation greater than 0.5 is $9/(16\pi f_{nd}) = 96$ symbols. A modified Jakes fading

model is used to model the time variations of each of the channel paths [40]. The pilots are positioned according to the optimal scheme found in the previous section ($\bar{L} = 1$). The number of basis functions for all simulated BE sets is

$$Q = 2\lceil N_{\text{bem}} f_{nd} \rceil + 1, \quad (38)$$

where $N_{\text{bem}} = 2N$. This number was tested numerically to enable good accuracy description of the channel with the BE complex exponents and polynomial functions (below 1% error). This is also the number of basis functions used in [30]. For the selection of eigenvectors as the functions set, we could have decreased this number slightly.

We present simulation results for the following equalization algorithms.

- (i) Maximum Likelihood equalization using perfect channel knowledge.
- (ii) Maximum likelihood equalization with channel estimation based only on the pilots. This is identical to the first iteration of the proposed algorithm.
- (iii) Time-varying BW algorithm with BE based on complex exponential functions (13) (BW-BE-exp).
- (iv) Time-varying BW algorithm with BE based on complex exponential functions (13) and initial channel guess identical to the true channel (BW-BE-exp & perfect init.). The difference between the error curve of this simulation and the previous one will indicate if we have an issue of convergence to a local maximum.
- (v) Time-varying BW algorithm with BE based on polynomial functions (14) (BW-BE-poly). This might be considered a significant improvement of [27].
- (vi) Time-varying BW algorithm with BE based on optimal basis functions (BW-BE-eig).
- (vii) Non-time-varying BW algorithm based on dividing the data blocks into shorter blocks in which channel is assumed to be constant (BW-SB). This is essentially the method of [23].
- (viii) The BW-RLS method in [26] (called APP-SDD-RLS in [26]). This method was initialized using the same initialization scheme we used for the BW-BE methods. After the parameters of the BE are found, the actual channel responded estimate is computed for every time instance. Finally, the BCJR algorithm uses this estimate to calculate the transitions probabilities which are the starting point for the BW-RLS in [26].
- (ix) Per-survivor processing with RLS channel estimator [8, 9].
- (x) Iterative Viterbi-based equalization with BE-based channel estimation. Reduced complexity variants of this algorithm appeared in [14, 18].

Simulation results for various signal-to-noise ratios (SNR) are presented in Figure 3 for block size of 256 symbols, 20 pilots (about 8% pilots), and multipath channel with three symbol-spaced paths with power profile $[0, -3, -3]$ dB. The proposed BE-based EM algorithm performance is very

close to the performance of a nonblind ML equalizer with perfect channel knowledge for both complex exponentials and polynomial BE functions. The BW-RLS method and the BW-SB have similar performance. This is not surprising as the BW-RLS method can be considered a non-time-varying BW that operates on a (exponentially weighted) sliding block whose length is proportional to the forgetting factor used. The BW-RLS can be therefore considered as a generalization of the BW-SB. The PSP-RLS method exhibits poor performance in these scenarios as the RLS component is not able to track the fast fading. The performance of Vit-BE is closer to our proposed methods although still significantly inferior. This is because the channel estimation part uses only hard decisions and does not weight the probability of different hypothesis for the transmitted sequence. The training based ML equalizer has the worst performance as the number of pilots is not sufficient for reliable channel estimation. It is therefore evident that the proposed semi-blind approach can reduce the number of required pilots significantly thereby increasing the available bandwidth for information.

Similar results were obtained for different channel profiles and delay spreads. For example, Figures 4 and 5 present the performance of the proposed algorithm for two equal power taps and four equal power taps, respectively. It is clearly shown that the proposed algorithm outperforms all previously proposed schemes that were simulated.

Results for block size 256 and SNR = 12 dB are shown in Figure 6 for various numbers of pilot symbols (the zero pilot percentage point in the figure corresponds to totally blind case and uses random initial guess). It is shown that in these conditions, our proposed method requires about 12% pilots to achieve close to optimal performance while Vit-BE requires up to 20% pilots to reach comparable performance and the other methods performance is even worse. Training-based equalization does not produce reliable results because the fast fading channel necessitates an excessive number of pilots for reliable estimation and tracking of the channel throughout the block in these conditions.

Results for various block sizes are shown in Figure 7. The performance of the proposed method improves with increasing the block size even though this means more functions are needed for modeling the time behavior of the channel. The reason is that the larger block size enables better filtering out of the noise as well as benefiting from more pilots in the block. All simulated methods reach a saturation point beyond which it is not beneficial to increase the block size further. The proposed method is shown to reach this point for smaller blocks thereby enabling latency reduction. From other simulation results (not shown here as they resemble the results presented so far), it is concluded that the proposed methods of BW-BE are superior over all simulated previously proposed methods in a large range of block sizes, pilot percentages, and channel profiles.

Finally, although a detailed convergence analysis is beyond the scope of this paper, a histogram of the number of iterations required to achieve convergence is plotted in Figure 8. This plot is for SNR = 12 dB and under the same conditions as in Figure 3. Convergence is determined based

on the Q function value, defined in (8). Convergence is assumed when the ratio of two Q function values of two consecutive iterations is sufficiently close to one, that is, less than $1 + 10^{-10}$. The maximum number of iterations allowed in this plot is 30. It is shown that about 75% of the blocks converge with 8 iterations or less and 90% of the blocks require 15 iterations or less for convergence. These results can be used to evaluate the computational complexity of the proposed algorithm.

7. Conclusion

A novel method for maximum likelihood semi-blind joint channel estimation and equalization for doubly selective channels is proposed. This method is based on expectation maximization (EM) principle combined with a BE model of the doubly selective channel. It is shown that the main drawback of the proposed approach is its possible convergence to local maxima. To alleviate this problem, we propose an initialization scheme to the equalization based on a small number of pilot symbols. We discuss the selection of basis functions set and find the optimal basis set. We further derive a pilot positioning scheme targeted to reduce the probability of convergence to a local maximum. The resulting algorithm is shown to be significantly superior over previously proposed equalization schemes and to perform in many cases close to an ML equalizer with perfect channel knowledge.

Appendices

A. Comparison of the proposed method to [27]

In this appendix, we highlight the differences between the algorithm proposed in [27] and the algorithm presented in this paper. For simplicity, we discuss only the non-time-varying case as the generalization to time-varying case is straightforward.

The channel estimation is updated in the BW algorithm at each iteration with:

$$\hat{\mathbf{h}}^{(l+1)} = \left(\sum_{n=0}^{N-1} \sum_{k=0}^{M^l-1} p(\mathbf{s}_{k,n} | \mathbf{y}, \theta^{(l)}) \mathbf{s}_{k,n}^H \mathbf{s}_{k,n} \right)^{-1} \times \sum_{n=0}^{N-1} \sum_{k=0}^{M^l-1} p(\mathbf{s}_{k,n} | \mathbf{y}, \theta^{(l)}) \mathbf{s}_{k,n}^H y_n. \quad (\text{A.1})$$

In [27], the channel is updated in two stages. First, the expectations of the Hidden Markov Model states are computed with

$$m_k = \frac{\sum_{n=0}^{N-1} p(s_{k,n} | \mathbf{y}, \theta^{(l)}) y_n}{\sum_{n=0}^{N-1} p(s_{k,n} | \mathbf{y}, \theta^{(l)})}, \quad (\text{A.2})$$

where m_k is the expected received value when in state k . In the second stage, the channel coefficients are updated by

fitting them to the values of $\mathbf{m} = [m_0, \dots, m_{M^L-1}]^T$ in a least squares sense, that is,

$$\hat{\mathbf{h}} = \arg \min_{\mathbf{h}} \|\mathbf{S}\mathbf{h} - \mathbf{m}\|^2, \quad (\text{A.3})$$

where \mathbf{S} is an $M^L \times L$ matrix whose rows are the states of the trellis, that is, $\mathbf{S} = [\mathbf{s}_0^T, \dots, \mathbf{s}_{M^L-1}^T]^T$. The well-known solution is

$$\mathbf{S}^H \hat{\mathbf{h}} = \mathbf{S}^H \mathbf{m}. \quad (\text{A.4})$$

Using the definition of \mathbf{S} and plugging (A.2) into (A.4), we obtain

$$\sum_{k=0}^{M^L-1} \mathbf{s}_k^H \mathbf{s}_k \hat{\mathbf{h}} = \sum_{k=0}^{M^L-1} \mathbf{s}_k^H \frac{\sum_{n=0}^{N-1} p(s_{k,n} | \mathbf{y}, \theta^{(l)}) y_n}{\sum_{n=0}^{N-1} p(s_{k,n} | \mathbf{y}, \theta^{(l)})}. \quad (\text{A.5})$$

Comparing (A.5) to (A.1), we see that the two methods are not equivalent. Moreover, the method in [27] converges to the true BW only when $\sum_{n=0}^{N-1} p(s_{k,n} | \mathbf{y}, \theta^{(l)})$ is constant for all values of k . This is likely to happen for long blocks. However, for short block it is not necessarily true and in this case the approach presented in [27] leads to a suboptimal solution.

B. Extension to correlated channel's paths

In this appendix, we extend the proposed equalization algorithm for the case of correlated channel's paths. First, let us assume for simplicity we have only two paths in the channel and rewrite (6) in the concatenated form

$$\begin{bmatrix} \mathbf{h}^{(0)} \\ \mathbf{h}^{(1)} \end{bmatrix} = \begin{bmatrix} B & 0 \\ 0 & B \end{bmatrix} \begin{bmatrix} \mathbf{g}_0 \\ \mathbf{g}_1 \end{bmatrix} = \bar{B} \bar{\mathbf{g}}. \quad (\text{B.1})$$

For the correlated case we can still have

$$\mathbf{h} = \bar{B} \bar{\mathbf{g}}, \quad (\text{B.2})$$

where now $\mathbf{h} = [\mathbf{h}^{(0)T}, \mathbf{h}^{(1)T}, \dots, \mathbf{h}^{(L-1)T}]^T$ and $\mathbf{h}^{(i)}$ is a vector of size N that represents the i th channel tap amplitude time variations. The size of the matrix B now is $LN \times Q$. The main difference in the correlated case compared to the uncorrelated case is that the matrix \bar{B} is no longer block diagonal and the off diagonal elements represent the correlation between the taps. In this description, there is a common coefficients set used for all channel taps and the difference between taps is due to different basis functions for each tap (the basis functions for the i th tap are the rows iN to $(i+1)N-1$ in the matrix \bar{B}). The criterion for basis selection, similar to (15), is

$$B = \arg \min_B E \|\mathbf{h} - \bar{B} \bar{\mathbf{g}}\|^2 \quad \text{s.t.} \quad \text{rank}[\bar{B}] = \bar{Q}, \quad (\text{B.3})$$

Obviously, \bar{Q} needs to satisfy $Q \leq \bar{Q} \leq QL$ and will vary according to how much the taps are actually correlated. In the uncorrelated case, we will have $\bar{Q} = QL$ and in fully correlated case $\bar{Q} = Q$. With uncorrelated taps, the autocorrelation matrix $E[\mathbf{h}\mathbf{h}^H]$ is block diagonal and there

are L blocks along the diagonal, each is equal to R_{corr} as defined in (5). The eigenvectors matrix in this case will also form a block diagonal matrix with L identical blocks whose first Q rows are the basis functions of the individual taps. In this case, the basis functions for each tap are identical and we return to the description of the channel response given in (9). In the correlated case, we have

$$\mathbf{h}_n = \bar{B}_n \mathbf{g}, \quad (\text{B.4})$$

where

$$\bar{B}_n = [\mathbf{b}_n^T, \mathbf{b}_{N+n}^T, \dots, \mathbf{b}_{(L-1)N+n}^T]^T, \quad (\text{B.5})$$

where \mathbf{b}_i is the i th row in the matrix \bar{B} . The matrix \bar{B}_n is then built from the L rows in B that correspond to time instant n . This description can then be used in the derivation of the EM procedure and the resulting updating equations are

$$\hat{\mathbf{g}}^{(l+1)} = \left(\sum_{n=0}^{N-1} \sum_{k=0}^{M^L-1} p(\mathbf{s}_{k,n} | \mathbf{y}, \theta^{(l)}) B_n^H \mathbf{s}_{k,n}^H \mathbf{s}_{k,n} B_n \right)^{-1} \quad (\text{B.6})$$

$$\times \sum_{n=0}^{N-1} \sum_{k=0}^{M^L-1} p(\mathbf{s}_{k,n} | \mathbf{y}, \theta^{(l)}) B_n^H \mathbf{s}_{k,n}^H y_n,$$

$$(\sigma^2)^{(l+1)} = \sum_{n=0}^{N-1} \sum_{k=0}^{M^L-1} p(\mathbf{s}_{k,n} | \mathbf{y}, \theta^{(l)}) |y_n - \mathbf{s}_{k,n} B_n \mathbf{g}|^2. \quad (\text{B.7})$$

C. Proof of the proposed BW initialization method

In this appendix, we prove the equivalence between the initialization method of the BW based on symbols priors, discussed in Section 3.3 and the initialization method shown in (21). First, define the group $\underline{n} = \{n, n-1, \dots, n-L+1\}$, that is, the time instances on which the transition $\mathbf{s}_{k,n}$ depends. In addition, define the group G as the group of all transitions in the trellis that are consistent with the pilots. Our branch metric initialization is

$$\gamma_{k,n} = \begin{cases} C, & \mathbf{s}_{k,n} \in G, \\ 0, & \text{otherwise.} \end{cases} \quad (\text{C.1})$$

The probability of a transition in the trellis is the sum of probabilities of all paths going through this transition. The probability of every path that is consistent with the pilots is C^N and

$$p(\mathbf{s}_{k,n}) = \begin{cases} M^{N-|\underline{n}|} C^N, & \mathbf{s}_{k,n} \in G, \\ 0, & \text{otherwise,} \end{cases} \quad (\text{C.2})$$

where $|\underline{x}|$ is the size of the group \underline{x} . We can now establish the identity

$$\begin{aligned} \sum_{k=0}^{M^L-1} p(\mathbf{s}_{k,n} | \mathbf{y}, \theta^{(l)}) \mathbf{s}_{k,n}^H \mathbf{s}_{k,n} &= C^N M^{N-|\mathbf{p} \cup \underline{a}|} \sum_{s_{n,k} \in G} \mathbf{s}_{k,n}^H \mathbf{s}_{k,n} \\ &= C^N M^{N-|\mathbf{p} \cup \underline{a}|} M^{L-|\mathbf{p} \cap \underline{a}|} \mathbf{z}_n^H \mathbf{z}_n \\ &= C^N M^{N-|\mathbf{p}|} \mathbf{z}_n^H \mathbf{z}_n, \end{aligned} \quad (\text{C.3})$$

$$= C^N M^{N-|\mathbf{p}|} \mathbf{z}_n^H \mathbf{z}_n, \quad (\text{C.4})$$

where $\mathbf{z}_n = [x_p(n), x_p(n-1), \dots, x_p(n-L+1)]$ and $x_p(n)$ is defined in (31). The identity (C.3) follows from the fact that the constellation has zero mean and (C.4) follows from $|\underline{a} \cup \underline{b}| = |\underline{a}| + |\underline{b}| - |\underline{a} \cap \underline{b}|$. Similarly, we have

$$\sum_{k=0}^{M^L-1} p(\mathbf{s}_{k,n} | \mathbf{y}, \theta^{(l)}) \mathbf{s}_{k,n}^H \mathbf{y}_n = C^N M^{N-|\mathbf{p}|} \mathbf{z}_n^H \mathbf{y}_n. \quad (\text{C.5})$$

Using (C.4) and (C.5), we can now show the equivalence. First, we have

$$\begin{aligned} \sum_{n=0}^{N-1} \sum_{k=0}^{M^L-1} p(\mathbf{s}_{k,n} | \mathbf{y}, \theta^{(l)}) B_n^H \mathbf{s}_{k,n}^H \mathbf{s}_{k,n} B_n \\ = \sum_{n=0}^{N-1} B_n^H \left(\sum_{k=0}^{M^L-1} p(\mathbf{s}_{k,n} | \mathbf{y}, \theta^{(l)}) \mathbf{s}_{k,n}^H \mathbf{s}_{k,n} \right) B_n \\ = C^N M^{N-|\mathbf{p}|} \sum_{n=0}^{N-1} B_n^H \mathbf{z}_n^H \mathbf{z}_n B_n \\ = C^N M^{N-|\mathbf{p}|} A_p^H A_p, \end{aligned} \quad (\text{C.6})$$

where the last identity follows directly from the definition of the matrices. Similarly, we have

$$\begin{aligned} \sum_{n=0}^{N-1} \sum_{k=0}^{M^L-1} p(\mathbf{s}_{k,n} | \mathbf{y}, \theta^{(l)}) B_n^H \mathbf{s}_{k,n}^H \mathbf{y}_n \\ = \sum_{n=0}^{N-1} B_n^H \left(\sum_{k=0}^{M^L-1} p(\mathbf{s}_{k,n} | \mathbf{y}, \theta^{(l)}) \mathbf{s}_{k,n}^H \mathbf{y}_n \right) \\ = C^N M^{N-|\mathbf{p}|} \sum_{n=0}^{N-1} B_n^H \mathbf{z}_n^H \mathbf{y}_n \\ = C^N M^{N-|\mathbf{p}|} A_p^H \mathbf{y}. \end{aligned} \quad (\text{C.7})$$

By combining the results from (C.6) and (C.7), it can be verified that (21) is identical to (B.6) for this particular choice of branch metrics.

Acknowledgment

Parts of the paper were presented at ISCCSP2008 [34].

References

- [1] O. Rousseaux, G. Leus, and M. Moonen, "Estimation and equalization of doubly selective channels using known symbol padding," *IEEE Transactions on Signal Processing*, vol. 54, no. 3, pp. 979–989, 2006.
- [2] G. B. Giannakis and C. Tepedelenlioglu, "Basis expansion models and diversity techniques for blind identification and equalization of time-varying channels," *Proceedings of the IEEE*, vol. 86, no. 10, pp. 1969–1986, 1998.
- [3] G. Leus, I. Barhumi, O. Rousseaux, and M. Moonen, "Direct semi-blind design of serial linear equalizers for doubly-selective channels," in *Proceedings of the IEEE International Conference on Communications*, vol. 5, pp. 2626–2630, Paris, France, June 2004.
- [4] S. Colonnese, G. Panci, S. Rinauro, and G. Scarano, "Semi-blind bussgang equalization for sparse channels," *IEEE Transactions on Signal Processing*, vol. 57, no. 12, pp. 4946–4952, 2009.
- [5] G. Leus, "Semi-blind channel estimation for rapidly time-varying channels," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, vol. 3, pp. 773–776, Philadelphia, Pa, USA, March 2005.
- [6] I. Barhumi and M. Moonen, "MLSE and MAP equalization for transmission over doubly selective channels," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 8, pp. 4120–4128, 2009.
- [7] R. Raheli, A. Polydoros, and C. Tzou, "Per-survivor processing: a general approach to MLSE in uncertain environments," *IEEE Transactions on Communications*, vol. 43, no. 2–4, pp. 354–364, 1995.
- [8] H. Zamiri-Jafarian and S. Pasupathy, "Adaptive MLSDE using the EM algorithm," *IEEE Transactions on Communications*, vol. 47, no. 8, pp. 1181–1193, 1999.
- [9] H. Kubo, K. Murakami, and T. Fujino, "Adaptive maximum-likelihood sequence estimator for fast time-varying intersymbol interference channels," *IEEE Transactions on Communications*, vol. 42, no. 2–4, pp. 1872–1880, 1994.
- [10] D. K. Borah and B. D. Hart, "Receiver structures for time-varying frequency-selective fading channels," *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 11, pp. 1863–1875, 1999.
- [11] Q. Dai and E. Shwedyk, "Detection of bandlimited signals over frequency selective Rayleigh fading channels," *IEEE Transactions on Communications*, vol. 42, no. 2–4, pp. 941–950, 1994.
- [12] R. A. Iltis, "A Bayesian maximum-likelihood sequence estimation algorithm for a priori unknown channels and symbol timing," *IEEE Journal on Selected Areas in Communications*, vol. 10, no. 3, pp. 579–588, 1992.
- [13] B. D. Hart and D. P. Taylor, "Maximum-likelihood synchronization, equalization, and sequence estimation for unknown time-varying frequency-selective channels," *IEEE Transactions on Communications*, vol. 46, no. 2, pp. 211–221, 1998.
- [14] S.-J. Hwang and P. Schniter, "Near-optimal noncoherent sequence detection for doubly dispersive channels," in *Proceedings of the Asilomar Conference on Signals, Systems, and Computers (ACSSC '06)*, pp. 134–138, Pacific Grove, Calif, USA, November 2006.
- [15] D. K. Borah and B. D. Hart, "Frequency-selective fading channel estimation with a polynomial time-varying channel model," *IEEE Transactions on Communications*, vol. 47, no. 6, pp. 862–873, 1999.
- [16] T. A. Thomas and F. W. Vook, "Multi-user frequency-domain channel identification, interference suppression, and

- equalization for time-varying broadband wireless communications,” in *Proceedings of the IEEE Sensor Array and Multichannel Signal Processing*, pp. 444–448, Boston, Mass, USA, March 2000.
- [17] T. Zemen and C. F. Mecklenbräuer, “Time-variant channel estimation using discrete prolate spheroidal sequences,” *IEEE Transactions on Signal Processing*, vol. 53, no. 9, pp. 3597–3607, 2005.
- [18] A. E.-S. El-Mahdy, “Adaptive channel estimation and equalization for rapidly mobile communication channels,” *IEEE Transactions on Communications*, vol. 52, no. 7, pp. 1126–1135, 2004.
- [19] E. Baccarelli and R. Cusani, “Combined channel estimation and data detection using soft statistics for frequency-selective fast-fading digital links,” *IEEE Transactions on Communications*, vol. 46, no. 4, pp. 424–427, 1998.
- [20] B. D. Hart and S. Pasupathy, “Innovations-based MAP detection for time-varying frequency-selective channels,” *IEEE Transactions on Communications*, vol. 48, no. 9, pp. 1507–1519, 2000.
- [21] L. M. Davis, I. B. Collings, and P. Hoehner, “Joint MAP equalization and channel estimation for frequency-selective and frequency-flat fast-fading channels,” *IEEE Transactions on Communications*, vol. 49, no. 12, pp. 2106–2114, 2001.
- [22] R. Otnes and M. Tüchler, “Iterative channel estimation for turbo equalization of time-varying frequency-selective channels,” *IEEE Transactions on Wireless Communications*, vol. 3, no. 6, pp. 1918–1923, 2004.
- [23] G. K. Kaleb and R. Vallet, “Joint parameter estimation and symbol detection for linear or nonlinear unknown channels,” *IEEE Transactions on Communications*, vol. 42, no. 7, pp. 2406–2413, 1994.
- [24] J. Choi, “A joint channel estimation and detection for frequency-domain equalization using an approximate EM algorithm,” *Signal Processing*, vol. 84, no. 5, pp. 865–880, 2004.
- [25] E. Aktas, “Belief propagation with Gaussian priors for pilot-assisted communication over fading ISI channels,” *IEEE Transactions on Wireless Communications*, vol. 8, no. 4, Article ID 4907469, pp. 2056–2066, 2009.
- [26] M. Nissilä and S. Pasupathy, “Adaptive Bayesian and EM-based detectors for frequency-selective fading channels,” *IEEE Transactions on Communications*, vol. 51, no. 8, pp. 1325–1336, 2003.
- [27] C. Antón-Haro, J. A. R. Fonollosa, C. Faulí, and J. R. Fonollosa, “On the inclusion of channel’s time dependence in a hidden Markov model for blind channel estimation,” *IEEE Transactions on Vehicular Technology*, vol. 50, no. 3, pp. 867–873, 2001.
- [28] S.-J. Hwang and P. Schniter, “Fast noncoherent decoding of block transmissions over doubly dispersive channels,” in *Proceedings of the 41st Asilomar Conference on Signals, Systems and Computers (ACSSC ’07)*, pp. 1005–1009, Pacific Grove, Calif, USA, November 2007.
- [29] S. Adireddy and L. Tong, “Optimal placement of known symbols for slowly varying frequency-selective channels,” *IEEE Transactions on Wireless Communications*, vol. 4, no. 4, pp. 1292–1296, 2005.
- [30] X. Ma, G. B. Giannakis, and S. Ohno, “Optimal training for block transmissions over doubly selective wireless fading channels,” *IEEE Transactions on Signal Processing*, vol. 51, no. 5, pp. 1351–1366, 2003.
- [31] A. P. Kannu and P. Schniter, “MSE-optimal training for linear time-varying channels,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP ’05)*, pp. 789–792, Philadelphia, Pa, USA, March 2005.
- [32] A. P. Kannu and P. Schniter, “Design and analysis of MMSE pilot-aided cyclic-prefixed block transmissions for doubly selective channels,” *IEEE Transactions on Signal Processing*, vol. 56, no. 3, pp. 1148–1160, 2008.
- [33] T. Whitworth, M. Ghogho, and D. McLernon, “Optimized training and basis expansion model parameters for doubly-selective channel estimation,” *IEEE Transactions on Wireless Communications*, vol. 8, no. 3, pp. 1490–1498, 2009.
- [34] G. Kutz and D. Raphaeli, “Maximum likelihood semi-blind equalization of doubly selective channels,” in *Proceedings of the 3rd International Symposium on Communications, Control, and Signal Processing (ISCCSP ’08)*, pp. 1580–1584, St. Julians, Malta, March 2008.
- [35] B. Sklar, “Rayleigh fading channels in mobile digital communication systems, part I: characterization,” *IEEE Communications Magazine*, vol. 35, no. 7, pp. 90–100, 1997.
- [36] L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv, “Optimal decoding of linear codes for minimizing symbol error rate,” *IEEE Transactions on Information Theory*, vol. 20, no. 2, pp. 284–287, 1974.
- [37] J. K. Tugnait and S. He, “Doubly-selective channel estimation using data-dependent superimposed training and exponential basis models,” *IEEE Transactions on Wireless Communications*, vol. 6, no. 11, pp. 3877–3883, 2007.
- [38] K. A. D. Teo and S. Ohno, “Optimal MMSE finite parameter model for doubly-selective channels,” in *Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM ’05)*, vol. 6, pp. 3503–3507, St. Louis, Mo, USA, December 2005.
- [39] M. Dong, L. Tong, and B. M. Sadler, “Optimal insertion of pilot symbols for transmissions over time-varying flat fading channels,” Tech. Rep. ACSP TR-01-03-02, Cornell University, January 2003.
- [40] Y. Li and Y. L. Guan, “Modified Jakes’ model for simulating multiple uncorrelated fading waveforms,” in *Proceedings of the IEEE International Conference on Communications (ICC ’00)*, pp. 46–49, June 2000.