Carrier Frequency Offset Estimation for Hiperlan/2

Yuan Zhang, Reza Hoshyar and Rahim Tafazolli

Centre for Communication Systems Research
University of Surrey
Guildford GU2 7XH
UNITED KINGDOM

***Abstract*: ETSI Hiperlan/2 is a European standard for the broadband wireless LAN based on OFDM modulation scheme. Generally, OFDM applications are very sensitive to the carrier frequency offset (CFO) and accurate frequency synchronization is demanded. In this paper, the effect of CFO in Hiperlan/2 is analyzed. Applicable algorithms for CFO estimation are proposed, and their performances are evaluated and compared under different channel conditions.**

I. INTRODUCTION

Hiperlan/2 and IEEE 802.11a represent two broadband Wireless LAN standards that operate in the 5 GHz band and choose Orthogonal Frequency Division Multiplexing (OFDM) as the modulation scheme for the physical layer. Hiperlan/2 standard for physical layer [1], has been approved by ETSI in February 2001. The transmission format on the physical layer is a burst, which consists of a preamble part and a data part. So rapid acquisition and synchronization are needed at the receiver.

The main synchronization tasks in such a TDD-TDMA mode network as Hiperlan/2 can be divided into three aspects, i.e. frame detection, symbol timing estimation and carrier frequency offset correction. The former two have been discussed in a paper [2] published by the Centre for Communication System Research and several algorithms been developed. In this paper, we will focus on CFO estimation and give the applicable algorithms for Hiperlan/2.

The paper is organized as follows: Section II gives the signal model and analyzes the effect of frequency offset on the system performance. In Section III, firstly the different preamble structures defined in the standard are introduced. Then algorithms using these preambles for CFO estimation are proposed and their performances are compared under different channel conditions. And some conclusions are followed in Section VI.

II. MODELING AND EFFECT OF CFO IN HIPERLAN/2

In Hiperlan/2, one OFDM symbol is constituted by a set of 52 subcarriers and transmitted with duration of $4.0\,\mu s$. There are 48 data subcarriers for data transmission and 4 pilot subcarriers for reference information. The length of the useful symbol part is equal to 64 samples in the time domain and cyclic prefix part is 16 samples. 64-IFFT are used to generate the baseband signals, and the baseband format of a transmitted symbol is given by

$$x_n = \frac{1}{\sqrt{N}} \sum_{k=-K}^{K} X_k e^{j2\pi kn/N};$$
$$n = 0,1,2,...N-1;\ N \geq 2K+1. \quad (1)$$

Here N=64, 2K=52. $\{X_k\}$ are complex data symbols transmitted through the k th subcarrier that can be the constellation of BPSK, QPSK, 16QAM or 64QAM. In order to facilitate implementation of pulse-shape filter, oversampling can also be applied by padding a number of zeros in the middle of the transmitted data and using N-IFFT, where N>64 and N= 2^n. One transmitted symbol is denoted as $\{x_{N-G},...,x_{N-1},x_0,...,x_{N-1}\}$, where the first G samples are cyclic prefix with duration 0.8 μs to eliminate ISI. The discrete format of the impulse response of the time-variant multipath channel is given by

$$h(n_\tau; n) = \sum_{p=0}^{P-1} h_p(n) \cdot \delta(n_\tau - \tau_p) \quad (2)$$

Here P denotes the number of paths and τ_p is discrete relative delay of the p th path.

Channel	Characteristic	RMS delay (ns)
A	Rayleigh	50
B	Rayleigh	100
C	Rayleigh	150
D	Rice	140
E	Rayleigh	250

Table 1: Hiperlan/2 Typical Channels

Usually there is some tolerance for symbol timing errors in OFDM, if the duration of the cyclic prefix greater than or equal to the time spread of the channel, which it is usually the case in Hiperlan/2 considering different channel conditions defined in the standard [3] (to see table 1). However, OFDM is extremely sensitive to CFO since any of CFO will degrade the orthogonality of subchannels and will cause ICI. Figure 1 depicts the effect of CFO on the BER performance at the receiver of Hiperlan/2 in AWGN channel. Differentially detected QPSK is employed without channel coding, interleaving and clipping effect, and at the same time, perfect symbol timing is assumed.

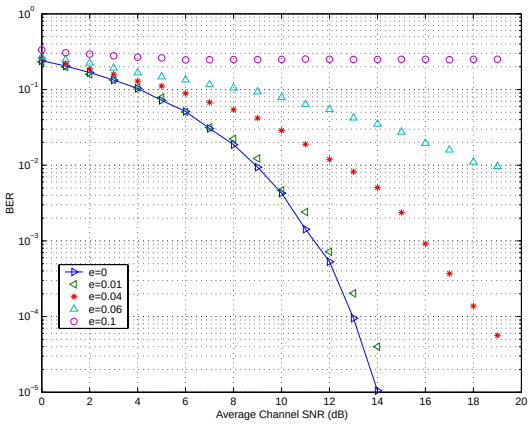


Figure 1 BER vs channel SNR performance over AWGN channels with CFO presented. Here e is the normalized CFO which is a fraction of subcarrier spacing.

From Figure 1 we notice that the receiver can tolerate only fractional CFO without greatly degrading system performance. For example, at SNR=14, for a negligible degradation of about 0.2 dB, the maximum tolerable frequency offset can not extend 0.01 of the subcarrier spacing.

The effect also can be analyzed mathematically. After passing through the fading channel, the received signal is sampled at the OFDM sample rate 1/T. The channel can be considered invariant during one symbol time, since coherence time of the channel is $T_c \approx 0.4 / f_{doppler} = 7.5ms >> 4.0\mu s$. Therefore, there is $h_p(n) = h_p$, $n = 0,1,2,...N-1$ in (2). Assuming perfect symbol timing the samples belonging to the first effective OFDM symbol can be written as

$$y_n = e^{j2\pi\varepsilon n/N} \cdot \sum_{p=0}^{P-1} h_p \cdot x_{n-\tau_p} + w_n \,, \qquad (3)$$

where ε is normalized frequency offset $\varepsilon = \Delta f / f_1$ (f_1 is subcarrier spacing), and w_n is the sampling output of AWGN. We note here if ε extends beyond subcarrier spacing, we can divide it into two parts: $\varepsilon = \varepsilon_F + \varepsilon_I$, where ε_I is the integral part of the frequency offset, and $0 \le \varepsilon_F < 1$ is the fractional part of the offset. The output at the m th subcarrier after FFT is

$$Y_m = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} y_n \cdot e^{-j2\pi mn/N} \qquad (4)$$

Substituting (1), (3) into (4) and rearranging yields,

$$Y_m = H_{m-\varepsilon_I} X_{m-\varepsilon_I} \frac{\sin(\pi\varepsilon_F)}{N\sin(\frac{\pi\varepsilon_F}{N})} \cdot e^{j(\frac{N-1}{N})\pi\varepsilon_F} + I_m + W_m \qquad (5)$$

where $H_k = \sum_{l=0}^{L-1} h_l \cdot e^{-j2\pi kn_l/N}$ is the transfer function of channel at the k th subcarrier; $I_m = \sum_{k \ne m-\varepsilon_I} H_k X_k (-1)^{k-m+\varepsilon_I} \frac{\sin(\pi\varepsilon_F)}{N\sin(\frac{\pi(k-m+\varepsilon_I+\varepsilon_F)}{N})} \cdot e^{-j\frac{\pi(k-m+\varepsilon_I)}{N}} \cdot e^{j\frac{\pi\varepsilon_F(N-1)}{N}}$ is ICI caused by other subcarriers; and $W_m = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} w_n e^{-j2\pi mn/N}$ is additive noise in the mth subcarrier bandwidth.

There are some remarks for Equation (5) as follows. The effect of fractional part ε_F and that of integral part ε_I are different on the demodulated symbols. The main effect of ε_I is to cause received subcarrier signals shift by ε_I subcarrier positions. Moreover, at the edge of bandwidth some signals modulated on the lowest and highest subcarriers are lost owing to this shift. On the other hand, ε_F causes a loss of mutual orthogonality between the subcarriers, then ICI occurs (to see I_m). Besides, the modulated symbol Y_m experiences an amplitude reduction and phase shift due to ε_F . The above two effects make the system performance very sensitive to ε_F , as has been depicted in Figure 1. Hence, accurate and applicable algorithms are needed for Hiperlan/2 to correct both integral CFO and fractional CFO, which is the topic of our further discussion in the following sections.

III. CFO ESTIMATION ALGORITHMS FOR HIPERLAN/2

In Hiperlan/2 physical layer standard [1], a maximum frequency offset of RF carriers at both Access Point (AP) and Mobile Terminal (MT) has been specified to be within ±20 ppm. Therefore, at a central frequency of 5.32 GHz, a maximum frequency offset of ±212.8 KHz could be experienced between any two transceivers, which is a fractional CFO about 0.68 of the subcarrier spacing. So an algorithm the correction range of which is at least of one subcarrier spacing is expected. It is meaningful to find an estimator with wider acquisition range for the following reasons. Considering the effect of Doppler shift added to the carrier inaccuracy, maybe the total CFO could be over one subcarrier spacing sometimes in the worst case. Moreover, the more CFO range we can estimate, the less accuracy of RF carrier is demanded. For example, if we can estimate a range of two subcarrier spacing (625KHz), a maximum frequency offset of about 60 ppm can be tolerated, which is a condition easily met by commercial devices.

There have been a lot of approaches to CFO estimation for OFDM in literature which can be categorized as data-aided methods [4]-[8] and blind ones [9]-[11]. Blind schemes usually exploit the specific redundancy associated with the cyclic prefix and suitable for the continuous transmission. Van de Beek gives a maximum likelihood estimator in ideal channels [9]. And some later

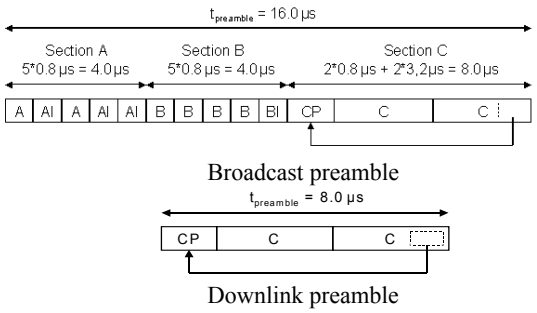
work extend it to multiuser systems [10], dispersive channels [11] respectively. Data-aided methods using reference symbols ahead of data payload are more suitable for fast and reliable synchronization required by burst transmission. A maximum likelihood estimator is given in [4], where two consecutive and identical reference symbols are used and CFO is calculated in frequency domain after taking the FFT. The acquisition range is limited to ±1/2 subcarrier spacing. Schmidl and Cox employ two training symbols to extend the estimation range to the overall transmitssion spectrum theoretically [5], where the first symbol is used to estimate fractional CFO and the second one for integral CFO. Modified versions of [5] are proposed in [6] and [7] based on different training symbol structures. Other approaches include Lambrette's scheme using single carrier training data [8], etc.

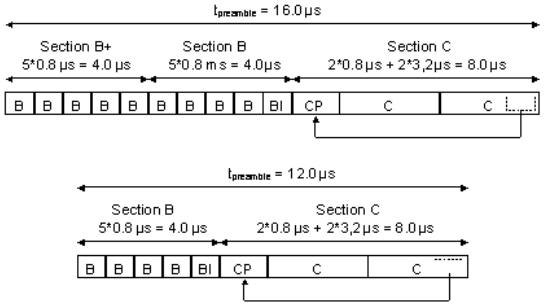
All the data-aided schemes mentioned above rely on their own defined training symbol structures, which are not definitely compatible with the Hiperlan/2 standard. Therefore, we need to modify these methods to adapt them to Hiperlan/2 standard and evaluate their performance in Hiperlan/2 simulation environment.

A. The Training Structures in Hiperlan/2

The Hiperlan/2 medium access control is based on a 2ms MAC frame sent continuously. The AP allocates time slots dynamically in the MAC frame for broadcast phase, downlink phase, uplink phase and random access phase. At the start of the frame is a broadcast burst, which informs MTs about time during the MAC frame in which they must listen to the transmitted data or they are allowed to transmit their data. According to the MAC protocol in Hiperlan/2, Five types of burst preambles have been defined in PHY layer standard [1], which can be used for time synchronization, frequency synchronization and channel estimation.

Figure 2 shows the structures of the preambles. A and B are short OFDM symbols of 16 time symbols, and AI/BI is the negative replica of A/B. Section A and Section B can be produced integrally by applying IFFT onto the frequency-domain sequences of 12 loaded subcarriers.





Long uplink preamble (direct link burst)/Short uplink preamble

Figure 2. Preamble structures in Hiperlan/2

C is long OFDM symbols with regular length, and CP is the cyclic prefix of 32 time symbols. In frgure 2, every type of burst preambles has the structure of Section B and Section C except downlink burst. It should be noted that in the structure of a MAC frame, the downlink burst is always preceded by a broadcast burst, and is received after the frequency offset. Therefore, the above two sections are used to implement CFO estimation algorithms.

B. Algorithms with Section B

Algorithm A: All the existing algorithms are based on the fact that CFO alters the phases of all subcarriers in exactly the same way, as for two identical parts of OFDM symbols in time domain, there is only a phase rotation between two parts if not considering the distortion caused by channel (to see (3)). Autocorrelation between these identical time domain samples (L samples long), therefore, can be implemented to estimate CFO. Correlation function is as follows,

$$P(d) = \sum_{k=0}^{L-1} y_{d+k+L} y_{d+k}^{*} \tag{6}$$

Then the phase rotation can be estimated as

$$\hat{\phi} == 2\pi\hat{\varepsilon}L / N = \arg(P(0)) \tag{7}$$

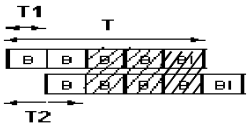
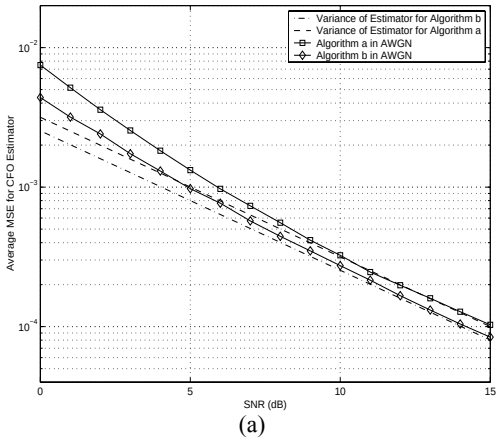


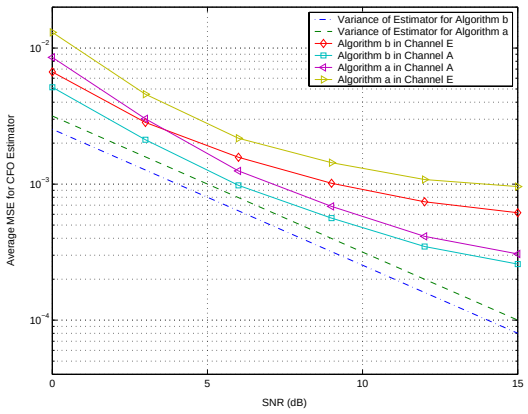
Figure 3. Description of Section B

The observation range of $\hat{\phi}$ is limited to $[-\pi,+\pi]$, so acquisition range of $\hat{\varepsilon}$ is limited to [-N/2L, +N/2L]. The shorter the length L of these identical parts, the larger estimation range that can be obtained, at the price of less accuracy since less samples are averaged.

In Section B, the preamble signal is periodic with both periods T_1 and T_2 (to see Figure 3), L can be chosen as T_1 and T_2 respectively. To get the maximum estimation range, T_1 is chosen.



(a)



(b)

Figure 4 MSE versus SNR in (a) AWGN (b) Channel A and Channel E

In order to add the average of the correlation, observation window can be extended to the shaded area for the reason that they rotate phases in the same mode. Then (6) is modified as

$$D(d)=\sum_{k=0}^{3L-1} y_{d+k+L}\,y_{d+k}^{*}+\sum_{k=3L}^{4L-1}(-y_{d+k+L})\,y_{d+k}^{*} \qquad (8)$$

Estimation of ε is given by

$$\hat{\varepsilon}=N\hat{\phi}/2\pi L=(2/\pi)\cdot\arg(D(0)) \qquad (9)$$

whose variance is calculated using the method given by [5]

$$\mathrm{var}(\hat{\varepsilon})=\frac{2}{\pi^2\cdot N\cdot SNR}=\frac{1}{32\pi^2\cdot SNR} \qquad (10)$$

Algorithm b: In [6], an algorithm is proposed which can be implemented in Hiperlan/2 using Section B with only slight modification. A training symbol with M (M>2) identical parts in time domain is assumed (in Section B, M=4). The proposed estimator exploits the correlations of the samples

$$R(m)=\sum_{k=mL}^{N-1} y_k\,y^{*}{}_{k-mL}\quad,0\le \mathrm{m}\le \mathrm{H} \qquad (11)$$

Here, L=N/M and H is a design parameter (to get the minimum variance, choose H=M/2). Considering the angles

$$\varphi(m)=[\arg\{R(m)\}-\arg\{R(m-1)\}]_{2\pi}\,,1\le m\le H$$

ε can be estimated by

$$\hat{\varepsilon}(m)=N\varphi(m)/2\pi L=2/\pi\cdot\varphi(m) \qquad (12)$$

Then the best linear unbiased estimator is given by

$$\hat{\varepsilon}=2/\pi\cdot(\frac{4}{5}\varphi(1)+\frac{1}{5}\varphi(2)) \qquad (13)$$

Here, var ($\hat{\varepsilon}$) achieves its minimum [6]

$$\mathrm{var}(\hat{\varepsilon})=\frac{3}{2\pi^2\cdot N(1-1/M^2)\cdot SNR}=\frac{1}{40\pi^2\cdot SNR} \qquad (14)$$

To evaluate the performances of the both algorithms, simulation has been run under AWGN, Channel A (indoor condition) and Channel E (large open space). The relative speed between AP and MT is considered as 5 km/hour, and at the same time, perfect timing is assumed. In Figure 4, the mean square error (MSE) of the estimation has been given as a function of average channel SNR, as well as the theoretical variance of the two algorithms. It can be seen that MSE in AWGN channel is quite close to the variance of the estimator especially for SNR>10. In multipath fading channels, performances of both algorithms experience a floor with increase of SNR, which is because of the ISI caused by the channel delay spread.

Algorithm a and *Algorithm b* are all unbiased estimators in nondispersive channel, and their acquisition range are all 2 subcarrier spacing. *Algorithm b* is superior to *Algorithm a* for MSE, which is 0.97 dB lower theoretically. However, it is at the price of greatly adding the computational complexity. *Algorithm a* needs 192 real products and 191 real additions, while *Algorithm b* requires 320 real products and 318 real additions. Considering the calculation load and performance, *Algorithm a* is more applicable.

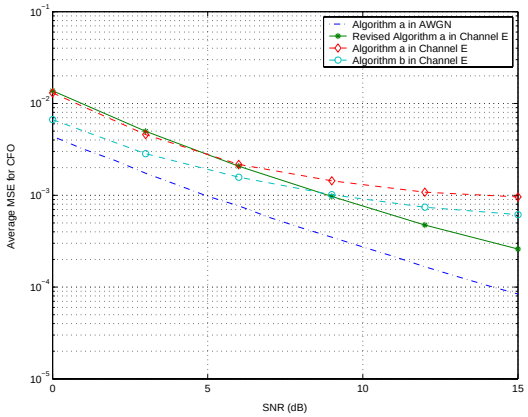


Figure 5 Performance of the *revised algorithm a*

Revised Algorithm a: In Figure 4 (b), it shows that performance of the estimators at high SNR in dispersive channels is much worse than that in AWGN. That is because in *Algorithm a* and *b*, the delay spread causes ISI and distorts the training symbols. To eliminate ISI, we

can consider the first identical part of Section B as the cyclic prefix and apply the correlation from the second part, that is

$$D(d) = \sum_{m=L}^{3L-1} y_{d+m+L} y_{d+m}^{*} + \sum_{m=3L}^{4L-1} (-y_{d+m+L}) y_{d+m}^{*} \qquad (15)$$

It is shown in Figure 5 that revised *algorithm a* significantly improves MSE performance at high SNR conditions in dispersive channels. Since CFO estimation usually is applied in a relative high SNR condition, *revised algorithm a* is superior to *algorithm a* and *b*, considering the performance and calculation load.

C. Algorithms with Section C

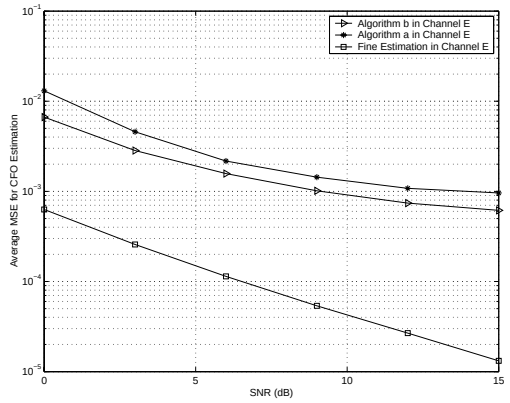


Figure 6 Performance of fine CFO estimation algorithm

Section C is formed with two identical symbols of regular length and a cyclic prefix of double length, which is designed for channel estimation. Section C cannot easily be used for initial estimation of CFO since the symbol length limits the estimation range. However, in some cases that more accurate estimation is needed, it can be applied for the fine CFO correction, in which the residual CFO is corrected after the coarse CFO estimation. Correlation Function (6) can be applied with L=N, d=32. The performance of fine CFO estimation is shown in Figure 6.

IV. CONCLUSIONS

In this paper, the frequency synchronization issues for Hiperlan/2 are studied. Sensitivity to CFO is analyzed both in mathematical aspect and simulation environment. It shows that only a fractional CFO not over 1% of subcarrier spacing can be tolerated without largely BER loss. Different algorithms for CFO estimation are presented, which are compatible with the preamble structures defined in Hiperlan/2 standard. It has been shown that *algorithm a* and *algorithm b* are two applicable methods for CFO initial estimation in Hiperlan/2, the acquisition range of which is large enough to satisfy the requirement. The *revised algorithm a* is the optimal method in dispersive channels and at high SNR level. And, the algorithm with Section C can be implemented for residual CFO estimation with satisfactory performance.

REFERENCES:

[1] Broadband Radio Access Networks; Hiperlan Type 2; Physical (PHY) Layer, Standard ETSI TS 101 475, ETSI, February 2001.

[2] Vicenc Almenar, Saied Abedi and Rahim Tafazolli, "Synchronization techniques for Hiperlan/2", in Proc. IEEE VTC Fall 2001, Vol. 2, pp. 762 –766, 2001

[3] J. Medbo, P. Schramm, "Channel Models for Hiperlan/2 in Different Indoor Scenarios", ETSI EP BRAN 3ERIO85B, March 1998.

[4] P. H Moose, "A Technique for Orthogonal Frequency Division Multiplexing Frequency Offset Correction", IEEE Trans. Commun., vol. 42, pp. 2908-2914, Oct. 1994.

[5] T. M Schimdl and Donald Cox, "Robust Frequency and Timing Synchronization for OFDM', IEEE Trans. Commun., vol. 45, pp 1613-1621, December 1997.

[6] M. Morelli and V. Mengali, "An improved frequency offset estimator for OFDM applications," IEEE Commun. Lett., vol. 3, pp. 75-77, Mar. 1999.

[7] Y. H. Kim, I. Song, S. Yoon and S. R. Park, "An Efficient Frequency Offset Estimator for OFDM Systems and Its Performance Characteristics," IEEE Trans. Veh. Technol., vol. 50, pp. 1307-1312, Sept. 2001

[8] U. Lambrette, M. Speth and H. Meyr, "OFDM Burst Frequency Synchronization by Single Carrier Training Data," IEEE Commun. Lett., vol. 1, pp. 46-48, Mar. 1997

[9] J. J. van de Book, M. Sandell and P. O. Borjesson, "ML estimation of Time and Frequency offset in OFDM systems." IEEE Trans. on Signal Processing, pp. 1800-1805, July 1997.

[10] J. J. van de Beek, P. O. Borjesson and S. K. Wilson, "A time and frequency synchronization scheme for multiuser OFDM," IEEE J. Select. Areas Commun., vol. 17, pp. 1900-1913, Nov. 1999

[11] S. Barbarossa, M. Pompili and B. Giannakis, "Channel-Independent Synchronization of Orthogonal Frequency Division Multiple Access Systems," IEEE J. Select. Areas Commun., vol. 20, pp. 474-486, February 2002

Enabling Fingerprint Authentication in Embedded Systems

P.S. Cheng, Y.S. Moon, Z.G. Cao, K.C. Chan, T.Y. Tang

Department of Computer Science and Engineering
The Chinese University of Hong Kong
Correspondence email: pscheng@cse.cuhk.edu.hk

Abstract

In this paper, we study different methodologies for implementing fingerprint authentication in embedded systems, namely the DSP and System-on-Chip approaches. Hardware experiments were conducted for evaluation of these approaches. Results show that the SoC system with fixed-point arithmetic support is probably the most ideal candidate.

1. Introduction

With the recent development in mobile commerce, authentication technologies quickly become critical parts in embedded devices like PDAs, cell-phones, etc. Among such technologies, fingerprint matching is the most common method. Yet, fingerprint matching algorithms involve substantial amounts of computation, making it a non-trivial task to achieve in the embedded devices. Several methods can be used to enable the implementation of fingerprint authentication in embedded systems. They include the Application Specific Integrated Circuit (ASIC), Digital Signal Processor (DSP) and System-on-Chip (SoC) methods. In this paper, we will review the fingerprint matching algorithms and study their implementation using DSP and SoC methods experimentally. Our experience should provide a guidance for future development of other biometrics applications in the embedded system environment.

2. Fingerprint embedded systems

Fingerprint system can be generally divided into two types, identification and verification applications.

Identification application means identification of a person from a group of individuals. It often searches a database for the identity of a person using the person's fingerprint as index. The database size can range from a few persons (home use) to hundreds of persons. The performance of such system is database size dependent. Access control is a typical identification example. Figure 1 shows the block diagram of a common access control system. Enrollment is conducted when a user first registered a fingerprint in the system. The core part of the system is the authentication device. It needs to handle different connection such as Ethernet, RS232 and Wiegand (an industrial standard security protocol sending bit strings of specified length and specified content) in order to communicate with enrollment machine, database server and security controller.

Since our concern is limited to embedded systems, we will focus on those applications involving only a few fingerprint records so that all the components in Figure 1 are hosed in one single embedded system and real time performance is expected.

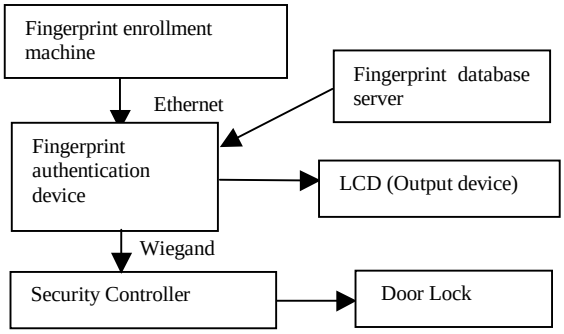


Figure 1: Block diagram of physical access control system

Fingerprint authentication often refers to the verification of a person's identity using his/her fingerprints. It is actually a 1-to-1 fingerprint matching problem. Such a problem can be well handled by a desktop PC in real time. The block diagram of fingerprint verification system is shown in Figure 2.

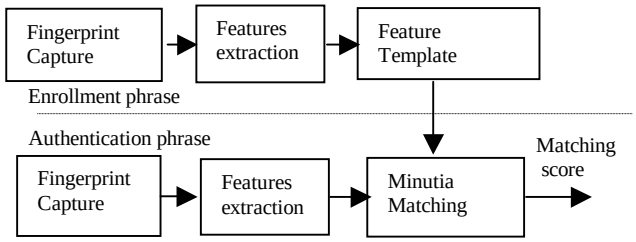


Figure 2: Fingerprint verification system

In mobile commerce activities, the authentication will be conducted in a PDA or cell-hone. When implementing such a system on an embedded device like a PDA, we encounter difficulties due to slower CPU speeds, absence of cache and most important of all, absence of a floating point unit. While optimizing the authentication, we cannot afford to sacrifice reliability. Thus, a fast, low cost and accurate methodology must be developed for fingerprint matching for embedded applications.

3. Fingerprint basics and system design

A human being's fingerprint contains numerous ridgelines. The end of a ridgeline forms a termination and the merge of two ridgelines forms a bifurcation as shown in Figure 3. A termination or a bifurcation is called a minutia point. The set

of the minutia points constitute the features characterizing a fingerprint. The matching of two fingerprints is based on comparing the existence and locations of the minutia points.

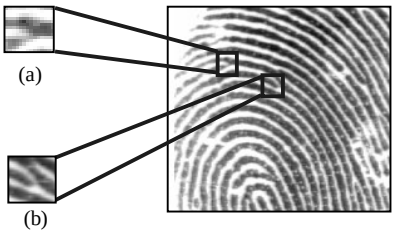


Figure 3: (a) Bifurcation (b) Termination

In this paper, our focus aims at accomplishing the fingerprint authentication in embedded processors in real time. As shown in the lower part of Figure 2, the authentication process is made of three steps. Fingerprint capture is basically an I/O process that can be accomplished in real time without difficulty. Minutia matching is point-matching technique. Its implementation is not complicated so that a smart card processor can be programmed to complete the job in real time [6]. The most time consumption processing is the feature extraction step. It also dominates the accuracy and performance of the authentication process.

Feature extraction (The direct gray scale approach) [7] can be divided into several parts. They are image enhancement, segmentation, image orientation, core point location and minutia extraction.

a) Image enhancement/filtering
   This part aims at enhancing the image quality before feature extraction. Common techniques like the application of a directional filter[8] or image normalization [9] is often employed.
b) Image segmentation and orientation
   This part computes the orientation of the fingerprint image and identifies the Region of Interest (ROI) [11].
c) Core point location
   This part locates a core point [10] for fingerprint alignment and matching. It is computed from the orientations of the ridgelines obtained from step (b).
d) Minutia extraction
   This part traces the ridgeline with the helps of orientation information and find out the bifurcation and termination features [14].

In general, the feature extraction process is a n^3 process, assuming that that fingerprint is a n by n image. Steps (a) to (d) involve intensive floating point calculations, especially approximations of trigonometric functions. Filtering can often be discarded in embedded system implementation of fingerprint verification systems.

4. Implementation

Like other embedded systems, the design of a fingerprint matching system requires the considerations of cost, development time and runtime performance. Since real time response is the critical factor in this case, we will focus on the searching of an appropriate implementation strategy that can yield such performance.

Previously, we had completed the development of a fingerprint authentication software on a PC. The 4000 lines program written in C language runs under the Microsoft Windows environment and can gives response in less than 1 second. When the same program was compiled to run in an embedded Linux environment under a PDA which has a 200 MHz StrongArm CPU, the average run time was more than 20 seconds. The obvious critical task is to enhance the program execution time. In this regard, we took two approaches to tackle the problem. They are DSP and SoC approaches.

4.1. The DSP approach

As the cost of the embedded fingerprint system should be as low as possible, we decided to conduct our work using a fixed point DSP rather than a floating point DSP [2]. For experiment purpose, we used a development board driven by a 40 MHz16-bit fixed point TI TMSC320C52 DSP with very limited on-chip data RAM and ROM. The TMS320C52 features a 4K*16-bit programmable ROM on-chip, and carries a 1056*16-bit data RAM on-chip. At the same time, the development board supports 32K*16-bit words of external memory. The DSP was connected to an Axis embedded Linux development board based 100 MHz embedded processor.

There are two methods to implement our fingerprint authentication program on the DSP platform:

1) Translate all the C source code to assembly language manually. This method gives a very high performance. But it is a very time consuming process.

2) Translate all the C source code to machine language under TI's C5X Code Composer Compiler [4].

To shorten the integration time, we decide to use the second method. With this approach, we can simulate the interface of software and hardware components in a short period of time. Although we use Code Composer as our development software, we still need to process certain modification when integrating to the DSP platform. First of all, the size of the fingerprint image to be processed has to be reduced to 64*64 pixels because of memory-limitation. Although it is harmful to the performance of the fingerprint authentication algorithm, our goal is to investigate the runtime of our algorithm. Secondly, our original program generates a lot of intermediate data. These data that were previously stored in dynamically allocated memory locations were reallocated to static memory locations since dynamic memory allocation is a very time consumption procedure in DSP processing.

Memory consumption is especially severe. Since the DSP and the host embedded processors operate at different speeds, it would not easy for them to share common memory. Moreover, the DSP accesses its built-in memory much faster than external memory. Therefore, fingerprint image data captured from a sensor must be transferred from the embedded

processor's memory to the DSP before the DSP can process them. In this way, memory requirement is doubled!

We have implemented both floating- and fixed-point versions of fingerprint authentication algorithm on the DSP. Figure 4 shows the number of instructions needed for the DSP to process the algorithm in the experiment.

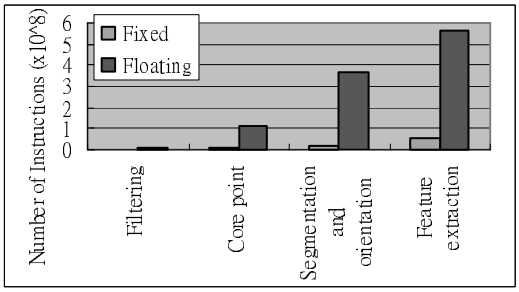


Figure 4: Comparison of number of instruction on TMS320C52 on fixed- and floating-point on fingerprint verification implementation

The total number of instructions needed for the fixed-point fingerprint algorithm is about 76 million while floating-point fingerprint algorithm needs about 1000 million. The floating-point version executes far more instructions because they are emulations only. The time measurement is about 1.9s and 26.1s on fixed-point and floating-point implementation respectively. Obviously, more time will become necessary to run the fingerprint authentication algorithm when the image size is 256*256.

Beside the DSP hardware, the embedded Linux system also contains extra hardware like the sensor, display, etc. The additional hardware connections increase the system complexity and reliability. This will increase the overall cost directly.

Therefore DSP design may not suitable for development of embedded fingerprint system.

4.2. SoC Implementation

Another approach for embedded fingerprint system is the use of System-on-Chip (SoC). SoC is a new type of hardware architecture with complex hardware that integrates different IP cores together. Besides the core processor is inside a SoC, it is also equipped with hardware controllers for different peripherals such as LCD and input/output device controllers. In fact, a SoC processor is more or less like the general-purpose processor but without floating point coprocessor because of cost and power consumption consideration. As all required processing have to be performed on the SoC as a standalone processor, this helps to reduce power consumption, minimize chip areas and simplify the hardware and software development process [3]. The ultimate question is to identify a suitable SoC processor and build an associated system capable of authenticating fingerprints in real time.

The use of SoC is suitable for embedded fingerprint verification system. First, as we have mentioned in Section 2, an embedded fingerprint system need to interface its users in the real world. SoC's built-in I/O controllers provide easy hardware interfaces to fingerprint sensors and LCD displays. With the use of SoC, system integration task shifts to software design.

Figure 5 shows the software architecture using the SoC design. Like a regular computer system, it can be divided into three layers, hardware layer, kernel layer and application layer [1]. With this approach, we can develop multiple applications on a single kernel for extensibility of the embedded fingerprint system. Besides, we can maximize the reusability on the software code. Nevertheless, we must verify if the layered approach will bring in an overhead that can delay the system response time.

In our experiment, we used the Intel StrongARM (206Mhz) embedded processor with Infineon capacitance fingerprint sensor in our system. The resolution of the sensor is 288x244 pixels with 8-bit gray scale image.

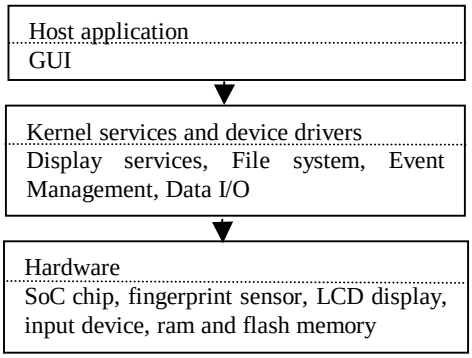


Figure 5: Software architecture on embedded fingerprint system design

Linking the host application and hardware is the Application Program Interface (API). We had chosen to use embedded Linux as the operating system because its open source code allows us to customize the software architecture for our system. With the use of Linux kernel, we can standardize the API for device drivers, scheduling and messaging services so as to save a lot of development time and effort.

For the application layer, we have to translate the fingerprint algorithm for StrongARM platform. As mentioned before, direct cross-complication of the algorithm is possible but resulted in poor runtime performance (>20s in our experiment). Therefore we started to develop a software reengineering process to port the software to its new environment. Basically, the reengineering process is made up of five steps:

(a) Whenever possible, replace floating-point data by integers;
(b) Replace the remaining floating-point calculations by fixed point calculations;
(c) Buffer all reusable complex calculations;
(d) Avoid the use of subroutines;
(e) Recompile using an optimized compiler.

The above procedure appears routine and straightforward. Yet, there are numerous domain specific conversions that must be taken care of. In our case, the precisions of the fixed-point system as well as good approximations of the trigonometric functions are our great concerns [12]. Care must be taken to ensure the reengineered software yields similar results as the original software. In our case, we estimated the errors by applying the reengineered software to a 300 persons database and analyzed the resultant False Acceptance Rates and False Rejection Rates to make sure that they resembled those of the original software. The 4000 lines of C codes take 2-3 months for reengineering. Ultimately, the authentication response is cut to one second or less.

For the development cost, porting the embedded Linux OS to StrongARM development board and writing drivers to interface the fingerprint sensors take 2 months to finish. Nevertheless, the experience and the source code are reusable.

5. Conclusions

To model the software development cost and time, we can use the Embedded Mode REVIC (Revised Intermediate COCOMO) Model [5], a cost modeling for embedded design system. It models the modern development technology such as re-use driven development, rapid development model, and application composition and generation capability [5]. In brief, the software development effort can be expressed in the following equation.

$$S_E = A * B * KSLOC^E * \prod_{i=1}^{15} F_i \qquad (1)$$

Where S_E is the software development effort

A is the constant used to compensate the effect of increasing project size

$KSLOC^E$ is the software size in thousands of source lines of code

B is the scale factor accounts for the relative economies or diseconomies of scale

F_i is the cost driver (multiplier) representing such project attributes like time constraints, reliability, software tools and application experience.

Optimization in DSP development process needs extensive programming experience in the operation of the DSP. This requirement increases cost driver F_i in equation (1) and hence, increases the development effort. When we switch to another type of DSP, utilization of the developed DSP code is very low because the algorithm for DSP is usually hardware dependent.

From the result, we conclude that the use of SoC is suitable for embedded fingerprint system development. It allows easy control of the peripherals and fast software reengineering time. This can achieve a fast development. Also, the simplicity of the hardware also enables manufacturers to embed fingerprint authentication systems in the future intelligent devices with low costs.

6. References

[1] Henry Chang, Larry Cooke, Merrill Hunt, Grant Martin, Andrew McNelly, Lee Todd, *Surviving the SOC Revolution, A Guide to Platform-Based Design*, Kluwer Academic Publishers, 1999, pp207-216

[2] Jennifer Eyre, *The Digital Signal Processor Derby*, Berkeley Design Technology Inc. IEEE Spectrum, Jun. 2001, pp.62-68

[3] Rolf Ernst, *Embedded System Architectures*, NATO Science Series: Applied Sciences, volume 357 of System-Level Synthesis, pages 1-43. Il Ciocco, August 1999.

[4] Pandey, R, *Advances in DSP Development Environments*, ELECTRO '96. Professional Program. Proceedings, 1996, pp299-301

[5] Sunita Devnani-Chulani, Brad Clark, Barry Boehm, *Calibration Approach and Results of the COCOMO II Post-Architecture Model*, Center for Software Engineering, University of Southern California

[6] Ho, H.C.; Moon, Y.S.; Ng, K.L.; Wan, S.F.; Wong, *Collaborative fingerprint authentication by smart card and a trusted host*, Electrical and Computer Engineering, 2000 Canadian Conference on, Volume: 1, 2000 Page(s): 108 -112 vol.1

[7] Maio, D.; Maltoni, D., *Direct gray-scale minutiae detection in fingerprints*, Pattern Analysis and Machine Intelligence, IEEE Transactions on, Volume: 19 Issue: 1, Jan 1997 Page(s): 27 –40

[8] Kamei, T.; Mizoguchi, M., *Image filter design for fingerprint enhancement*, Computer Vision, 1995. Proceedings., International Symposium on , 21-23 Nov 1995 Page(s): 109 –114

[9] Jain, A.; Lin Hong; Yifei Wan**,** *Fingerprint image enhancement: algorithm and performance evaluation*, Pattern Analysis and Machine Intelligence, IEEE Transactions on, Volume: 20 Issue: 8, Aug 1998 Page(s): 777 –789

[10] Asker M. Bazen and Sabih H. Gerez, *Extraction of Singular Points from Directional Fields of Fingerprints*, Mobile Communications in Perspective, Annual CTIT Workshop, Enschede, The Netherlands, February 2001

[11] Jain, A.; Lin Hong; Bolle, R., *On-line fingerprint verification*, Pattern Analysis and Machine Intelligence, IEEE Transactions on, Volume: 19 Issue: 4, April 1997 Page(s): 302 –314

[12] Moon Y.S., Luk T.C., Tang T.Y., Chan K.C., *The fixed-point implementation of fingerprint minutiae extraction algorithm*, manuscript in preparation.

Efficient Multidimensional QoS-Based Packet Scheduling for Mixed Services in HSDPA System

Saied Abedi, Sunil Vadgama
Fujitsu Laboratories of Europe LTD. (FLE)
Hayes Park Central, Hayes End Road, Hayes
Middlesex, UB4 8FE U.K.
Contact Email: S.Abedi@fle.fujitsu.com

Abstract

In High Speed Downlink Packet Access (HSDPA) system [1], it is highly desirable for real time conversational and non-real time services to efficiently share the available channels and bandwidth. In HSDPA system based on the reported channel state information from User Equipment (UE) and the other available data, Node B or base station performs a packet scheduling and channel assignment process per each Transmission Time Interval (TTI) to decide about the format of transmission on available parallel downlink channels. We have proposed a novel Multidimensional QoS-based Packet Scheduler (MQPS) for HSDPA system which in a mixed and single service environment outperforms Max C/I scheduler in terms of delay of the packet scheduling process whilst approaching the performance of Max C/I in terms of sector throughput bit rate. Taking into account all the aspects of QoS provisioning simultaneously, the proposed MQPS in this paper by reaching almost better level of fairness than PF scheduler, outperforms it in term of average delivered user throughput, regardless of the distance of the UE from the Node B. It also outperforms PF in terms of delivery delay of packets. In a mixed service environment in order to achieve a high packet scheduling performance in terms of delay profile and achieved throughput, by avoiding the multiple fixed defined schedulers, the proposed unified multidimensional fast dynamic packet scheduler elegantly and efficiently encapsulates features of many of possible packet scheduling strategies.

Keywords

Wireless Access, Scheduling, Multimedia

I. INTRODUCTION

Wireless multimedia applications are essential parts of enhanced 3G systems. High Speed Downlink Packet Access (HSDPA) is a promising technology that enhances the throughput and Quality of Service (QoS) of the third generation communication systems significantly. In HSDPA system number of parallel shared channels and higher levels of Modulation and Coding scheme (MCS) are employed to achieve a high data rate transfer from Node B or base station to User Equipment (UE). In order to select appropriate MCS level in the HSDPA system, each UE estimates the channel quality and reports the estimated Carrier-to-Interference Ratios (C/I values) to Node B [1]. Based on these reported values and the other available data, Node B performs a channel assignment and packet scheduling process for active users. Variety of packet scheduling techniques is proposed in literature [2-7].

Packet scheduling techniques such as Proportional Fair (PF) [6] or Max C/I satisfy only some aspects of QoS provisioning. PF scheduler tries to improve the fairness of packet scheduling and channel assignment by providing almost same level of average user throughput for UEs irrespective of their distances from Node B [4]. PF does not necessarily provide a good overall system throughput. The delay profile of delivery process is also sacrificed by PF which provides a poor overall delay profile comparing to the techniques such as Max C/I. The delay and jitter becomes highly important for real-time applications.

In this paper we propose a novel Multidimensional QoS-based Packet Scheduler (MQPS) for HSDPA system which outperforms Max C/I scheduler in terms of delivery time of the packet scheduling process whilst approaching the performance of Max C/I in terms of sector throughput bit rate. Taking into account all the aspects of QoS provisioning simultaneously, the proposed MQPS reaches almost better level of fairness than PF scheduler and outperforms it in terms of average user throughput delivered, regardless of the position of the UE in the Node B. It also outperforms PF in terms of delivery delay of packets.

Scheduling techniques such as PF could provide a fair output for the wireless end-users over a long period of time (as $time \to \infty$) [6],[7]. However the proposed MQPS provides the most possible equal outcome for wireless end-user under short periods of time with maximum achievable throughput within the assigned delay tolerance threshold and with maximum possible average throughput.

Under a mixed real time and non-real time mixed service scenario, existing packet scheduling and radio resource management techniques employ service classification and partitioned resource shaping [11]. Multiple schedulers are also used to respond to different requirements of individual classes of services. The partitioned

resource shaping with either fixed or a slow changing dynamic proves difficult and inefficient under fast changing dynamics of radio channel. While resources and schedulers assigned to one partition is under a high pressure to provide the demanded Quality of Service (QoS) requirements, the schedulers and resources assigned to other partitions are underutilized. Moreover efficiency and performance of such schedulers degrade considerably as the size of the shared bandwidth and the range of services having differing requirements increase. Attempts to recover some of the lost efficiency in such schedulers only ends in increased computational complexity and resulting costs.

In contrast, our proposed MQPS employs a unified packet scheduling technique for mixed traffic handling in HSDPA system which avoids such fixed or slow dynamic partitioned resource shaping. The proposed technique performs a fast (max) dynamic "on the fly" (TTI (Transmission Time Interval) by TTI) multi-dimensional resource monitoring, analysis and resource allocation. The proposed technique handles the mixed services in a way to satisfy all aspects of QoS provisioning such as priority, fairness, delivery delay, maximum achievable bit-rate and throughput. By monitoring the channel conditions, current data congestion, priority of services, delay and QoS profiles of existing users, it can transform itself to the best possible packet scheduling strategy by activating the relevant and appropriate features inside it. This process is performed on the fly and per TTI basis to achieve the highest possible level of packet scheduling performance. Through detailed simulation studies the overall performance of HSDPA system employing different packet scheduling techniques in terms of throughput, delay and fairness of packet delivery process is evaluated. It is shown that the proposed MQPS is able to outperform existing wireless packet scheduling techniques with considerable performance gain.

II. Packet Scheduling Techniques for HSDPA Systems

In the HSDPA system, in order to use all the free channels during the reporting period more efficiently, an asynchronous N-Channel Stop-and-Wait (SAW) Hybrid ARQ protocol is employed. This protocol gives the users ability to occupy the existing idle channels without waiting for the Acknowledge and Non-Acknowledge (Ack/Nack) messages of the previously transmitted packets.

In order to increase bit rate and efficiency of the system, a number of parallel channelisation codes in Node B is employed. To assign these channelisation codes to the best combination of the UEs, it is important to adopt a robust radio resource management and channel assignment policy.

Max C/I scheduler is one of the well-known scheduling techniques that first ranks the UEs in terms of their channel quality. UE with the best C/I reported value has the highest rank. Then the channels are exhaustively allocated starting from the users with best channel conditions. Max C/I scheduler dedicates itself to the overall delivered throughput without dealing with fairness issue efficiently. Consequently the UEs that are located far away from Node B receive less data packets and bad delay profiles. Consequently the UEs located far away from Node B experience poor quality of service than the UEs near to the base station.

Proportional Fair (PF) is another packet scheduling technique [6], [7], [4] which provides a better fairness than Max C/I scheduler. To rank the UEs first a fitness or credit value is assigned to each UE, for example as a ratio of instantaneous C/I reported value to long-term averaged SIR value.

This fitness value is used to rank users in terms of their eligibility for transmission. The shared channels are then allocated exhaustively among users. One possibility for allocation of channels is to allocate the available channel codes based on the normalized amount of data waiting for delivery in First-in First–out (FIFO) at Node B for each UE. Starting with user with highest fitness the channels are allocated based on normalized FIFO length. User with the higher amount of normalized queue length gets more number of channel codes. In this paper this approach for PF is adapted to allocate the channels amongst the candidate ranked UEs.

III. Multidimensional QoS-Based Packet Scheduler

The proposed MQPS first performs a *global ranking* of the existing User Equipment (UEs). Then in order to divide the available channels and resources between the ranked UEs, it performs a global resource allocation (*global weighting*) process. Both global ranking and weighting processes first build a profile for each UE based on all the dimensions involved in an efficient packet scheduling process. These parameters which are being monitored per TTI include: (a) QoS based metric which deals with the proportion of throughput of each user delivered within a QoS related delay tolerance threshold. Better this ratio, lower the rank and the chance to get more channels. (b) Dimension related to Carrier-to-Interference Ratio and channel conditions. (c) Dimension related to the estimated deliverable packet data units (octets) per UE. (d) Fairness related metric which measures the undelivered proportion of throughput per TTI for each UE. Higher this proportion betters the ranking and the chance to get more channels. (e) The dimension related to the delay of the first waiting packet or data unit in the FIFO queue in Node B to be delivered and the distance of this delay from delay tolerance threshold of service assigned to UE. Other dimensions such as impact of each user or UE on profit of operator considering the priority of UE or a dimension related to impact of each user or UE

on cost and revenue of the operator considering the priority of UE can be considered. The latter two metrics are not included in this paper. Global ranking and weighting processes by applying a linear or non-linear mapping, first maps the range of variations of these dimensions to a unified range. The final unified and weighted values are the fitness values which are employed to carry out the ranking and the exhaustive channel allocation process.

IV. All-IP Wireless Multimedia System Model Employing the HSDPA

It is assumed that a base station employing HSDPA scheme is serving a number of UEs, some with real-time video sessions and others with WWW browsing sessions in a wireless environment. The system model is depicted in Figure 1. A video traffic model based on H.263 video coding [12], [13] is applied. The model concentrates on the traffic characteristics related to video encoding and RTP (Real Time Protocol) packet transport.

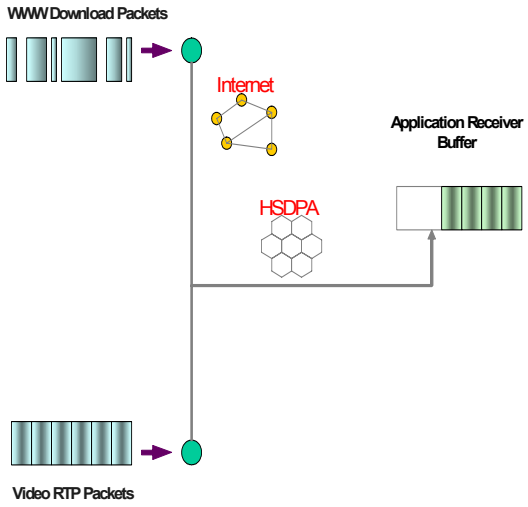


Figure 1 Realization of an All-IP wireless multimedia system employing HSDPA system

V. Simulation Results and Conclusions, Single WWW Traffic

We have developed a system level simulator for HSDPA system based on the proposed conditions in [1]. First we consider a single services scenario. Number of WWW browsing sessions is considered which consist of a sequence of packet calls. UEs are allocated in cells on uniform basis and then move around. Adjacent cell interference is the results of transmissions from the adjacent Node Bs. The inter-site distance is assumed to be 6 km. The path loss is considered to be present and affect the signal quality. To model the impact of Rayleigh fading, ETSI 6-path Rayleigh Vehicular A channel is employed [8]. It is assumed that the UEs' speed is 3.6 km/h. The shadowing is assumed to have a log-normal distribution. The decorrelation distance is 50 m. The difference between the arrival time and successful delivery to UE is considered as delivery delay. Minimum reporting delay is considered to be 3 TTIs (i.e.: 3 x 2ms). The average user throughput and service throughput metrics are defined in [4]. The simulation period is 60 sec or 30000 TTIs. A packet is dropped, if it can not be delivered within six retransmissions [1]. No limitation from higher layer is applied on the delivery delay. Therefore the exact value of delivery delay is being monitored for each transmitted data unit. In Figure 2 the performance of the proposed scheduler is compared to Max C/I and PF in terms of average user throughput. For each set of results a fitted curve has been determined and shown in Figure 2 along side with the original data. We define the average user throughput in terms of bits per second as average of ratios of number of bits in each packet call to the time required to transmit these bits [4].

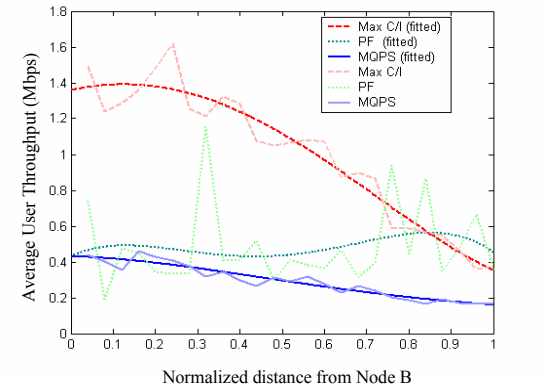


Figure 2 Average throughput of UEs versus normalized distance from Node B, fitted curve, 150 UEs per cell, each data set is accompanied by its correspondence fitted curves

We define the 95 percentile delay as the delivery delay within which 95% of data packets were successfully delivered [5]. In Figure 3 and 4 it is shown that MQPS provides a better average delivery delay and 95 percentile delay profile than PF and Max C/I.

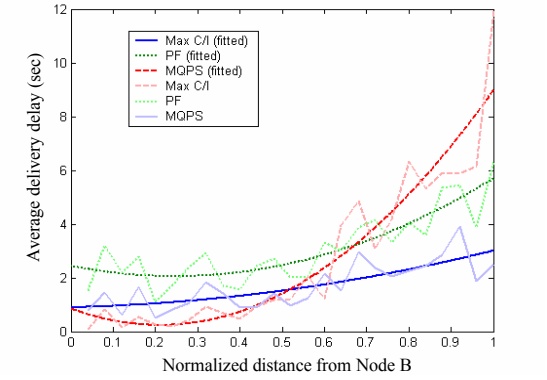


Figure 3 Average delivery delay versus normalized distance from Node B, original data and fitted curves

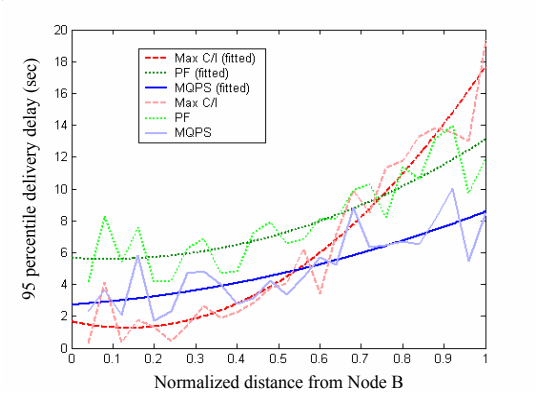


Figure 4 The 95 Percentile delivery delay versus normalized distance from Node B, original data and fitted curves

In an alternative definition of throughput, we define individual throughput as the ratio of successfully delivered data over the arrived data in Node B for each TTI. The real-time fairness of packet scheduling process is related to instantaneous variance of vector of individual throughputs. Lower this variance, better the fairness of the packet scheduling process. In Figure 5 real-time monitored variance is depicted.

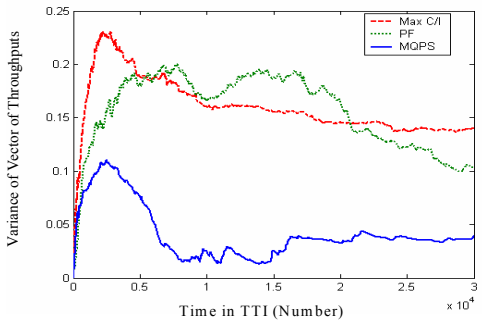


Figure 5 Real Time variations of variance of vector of individual throughputs, lower variance is equivalent to better fairness

Despite showing slightly less user throughput than PF in Figure 2, MQPS has managed to perform a faster convergence to an equal outcome to wireless end users (low variance) in Figure 5. PF does show a slow convergence time. Therefore MQPS has managed to outperform PF in terms of real-time fairness of delivery, all over the transmission period, providing better average and percentile delay profiles.

VI. Simulation Results and Conclusions, Mixed Traffic, Real-Time Video and WWW Traffic

Under the similar conditions as previous experiment and error-free channel estimation, mixed service scenario is considered. Only the inter-site distance is reduced to 2.8 km. For video session the delay tolerance threshold is assumed to be 100 ms [8], [9]. It is assumed that the video frame rate is 7.5 frames/sec and the target bit rate of the output video stream is 32 kbps. Initially 20 real-time video sessions and 30 WWW download sessions are present. In Figure 6 and 7 it is shown that MQPS provides a better average delivery delay for both video and WWW users.

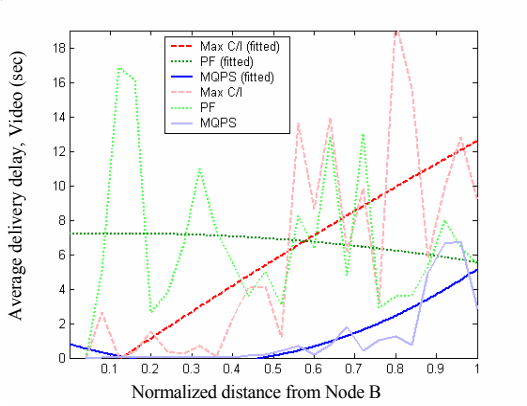


Figure 6 Average delivery delay for video sessions versus normalized distance from Node B, original data and fitted curves

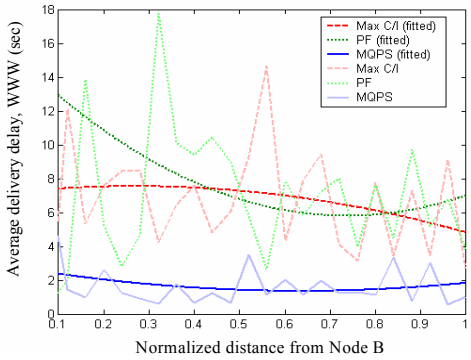


Figure 7 Average delivery delay of WWW users versus normalized distance from Node B, original data and fitted curves

Figure 8 and 9 show the 95 percentile delays Vs normalized distance between UE and Node B for WWW users and video users respectively. It can be seen that for both WWW and video users MQPS delivers better 95 percentile delay than Max C/I and PF.

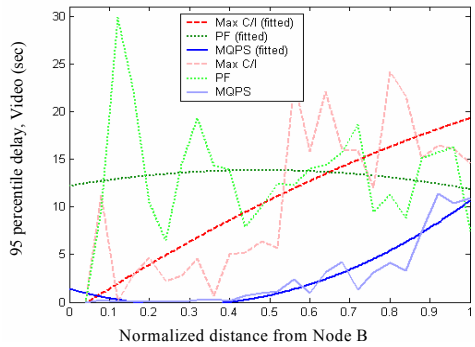


Figure 8 The 95 percentile delivery delay for video sessions versus normalized distance from Node B, original data and fitted curves

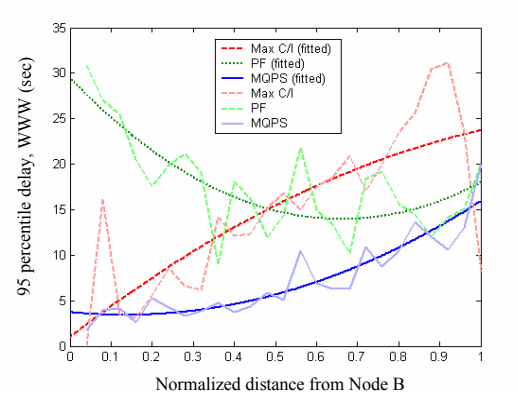


Figure 9 The 95 percentile delivery delay for WWW sessions versus normalized distance from Node B, original data and fitted curves

Superiority of MQPS over PF and Max C/I packet schedulers in terms of real-time fairness of packet delivery process is depicted in Figure 10. It can be seen that under current heavily loaded scenario MQPS has managed to provide a faster convergence than the other two techniques to a very low variance of vector of individual throughputs. The presented results clearly highlight the advantages of the proposed MQPS in this paper. Employing MQPS in Node B for HSDPA and future all-IP wireless systems will lead to significant and simultaneous improvements in terms of fairness, delay, individual user throughput providing a good overall service throughput and a better QoS.

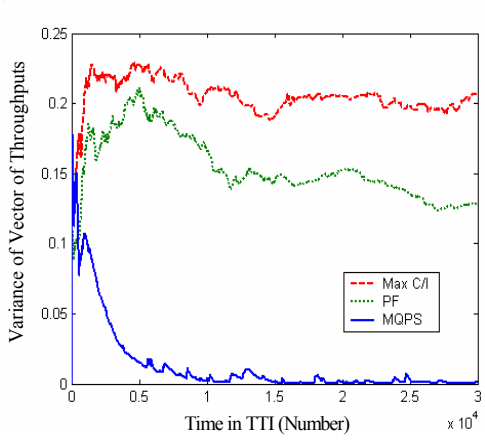


Figure 10 Real-time variations of variance of vector of individual throughputs (ratios), lower variance is equivalent to better fairness, fast convergence of MQPS to instantaneous equal output for wireless end-users

In Figure 11 performance in terms of average user throughput versus the normalized distance of UEs from Node B is depicted for video sessions. It can be seen that MQPS outperforms both the Max C/I and PF scheduler algorithms in terms of individual user bit-rate and distance related fairness.

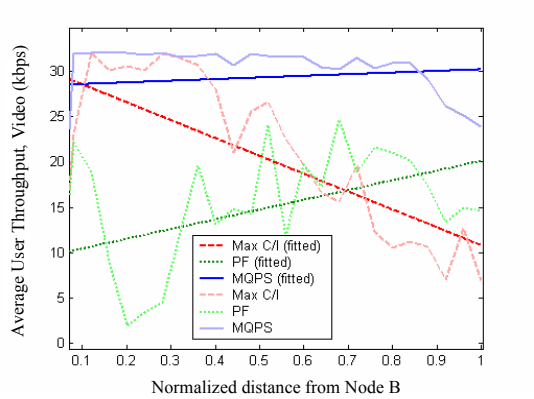


Figure 11 Average throughputs for video sessions versus normalized distance from Node B, original data and fitted curves

The presented results clearly highlight the advantages of the proposed MQPS in this paper. Employing MQPS in Node B for HSDPA and future all-IP wireless systems will lead to significant and simultaneous improvements in terms of fairness, delay, individual user throughput providing a good overall service throughput and a better QoS.

VII. REFERENCES

[1] TR25.848, ver1.0.0, RP-010191, TSG-RAN#11, March 2001.

[2] S. Abedi, S. Vadgama, "Hybrid Genetic Packet Scheduling and radio Resource Management for High Speed Downlink Packet Access", WPMC 2002 Conference, Hawaii, pp. 1192-1196

[3] S. Abedi, S. Vadgama, "A Radio Aware Random Iterative Scheduling Technique for High Speed Downlink Packet Access", VTC 2002, Fall, vol. 4, pp. 2322 – 2326, 24-28, Sept. 2002

[4] Y. Ofuji, S. Abeta, A. Morimoto, M. Swahashi "Comparison of Packet Scheduling Methods Focusing on Throughput of Each User in High Speed Downlink Packet Access", Technical Report of IEICE, MoMuC2002-3, (2002-05), pp. 51-58

[5] 3GPP TSG RAN WG1#21, Philips Technical Document, Turin, Italy, 27-31 August 2001.

[6] R. Agrawal, et. al. "Class and Channel Condition Based Scheduler for EDGE/GPRS", Proc. of SPIE, vol.#4531, Aug 2001

[7] A. Pandey, S. Emeott, J. Pautler, K. Rohani, "Application of MIMO and Proportional Fair Scheduling to CDMA Downlink Packet Data Channels", VTC 2002, Fall, vol. 2, pp. 1046 – 1050, 24-28, Sept. 2002

[8] ETSI/TR 101 112 v3.2.0(1998-04) "Selection procedures for the choice of radio transmission technologies of the UMTS".

[9] ARIB TR-T63-22.105, C3.10.0, “Services and Service Capabilities”, (3G TS 22.105 version 3.10.0 Release 1999), October 2001

[10] 3GPP TS 23.107, v5.1.0 “QoS Concept and Architecture” (Release 5), June 2001

[11] M. Ono, Y. Matsunaga, M. Momona, K. Okanoue, “A Proposal of All-IP Mobile Wireless Network Architecture QoS Packet Scheduler for Base Stations”, Technical Report of IEICE, MoMuC2002-3, (2002-05), pp. 13-18

[12] ITU-T Recommendation H.263, “Video Coding for Low Bit Rate Communication”, 02/1998

[13] H. Nyberg, C. Johansson, B. Olin, “A Streaming Video Traffic Model for the Mobile Access Network”, VTC 2001 Fall, IEEE VTS 54th, Vol. 1, pp. 423 –427.

Convergence Improvement of an Iterative Linear Multiuser Detector for DS-CDMA Systems

M. Mozaffaripour, R. Tafazolli
Mobile Communications Research Group
Centre for Communication Systems Research (CCSR)
University of Surrey, Guildford, Surrey, GU2 7XH, United Kingdom
Tel: +44-1483-689489;Fax: +44-1483-686011
M.Mozaffaripour@surrey.ac.uk
R.Tafazolli@Surrey.ac.uk

Abstract- This paper considers an iterative implementation of linear multiuser detection based on a Taylor series expansion of the correlation matrix. This method is less complex and easy to implement, yet its convergence rate depends on the eigenvalues of the correlation matrix. By analytical examination of the structure of this matrix, a simple method has been derived for calculating the proper value for the desired parameters. It works well in synchronous systems but its performance is not convincing in asynchronous ones. Using an algebraic theorem, better estimation of the parameters can be achieved that leads to better performance in the asynchronous scenarios. The performance of the latter method is also compared to the partial parallel interference (PPIC) through the simulations.

Index Terms- Multiuser detection, linear detectors, Decorrelator, Gershgorin Theorem, Matrix inversion

I. INTRODUCTION

Code Division Multiple Access (CDMA) offers many attractive properties as an access scheme for mobile communications. However, receivers using conventional detectors suffer from Multiple Access Interference (MAI). The system capacity and performance is degraded when the number of users increase and when the system operates in severe near-far environments. These conditions are inherent in cellular systems. Since the work of Verdu [2] on the optimum detector, several sub-optimum multi-user detectors have been proposed for improving capacity and mitigating the MAI problem of the conventional method. Amongst them are linear detectors of which Decorrelator and MMSE (Minimum Mean Square Error) are well-known approaches. These methods have several attractive properties but suffer from implementation complexity, which is mainly due to a need to invert a large matrix. Some work has been done to approximate the inverse matrix without computing directly [1,3]. For example [3] uses polynomial expansion (PE) method. This method is powerful, however it needs to calculate a set of coefficients. On the other hand, [1] has proposed a simplified polynomial-expansion, for which the convergence rate is rather slow because of the inaccuracies. Our aim is examining a modified version of a simple polynomial expansion method and achieving a new accurate set of convergence conditions.

In this paper we first briefly review the system model of CDMA and linear detectors in section II. In section III an iterative implementation of the Decorrelator is addressed. In section IV, based on analytical analysis, an accurate, yet simple method to estimate the parameters involved is proposed. Section V considers the Gershgorin algorithm, which improves the performance even in the asynchronous environments, as it estimates the parameters involved more accurately. The results have come in section VI. In section VII there will be the conclusions.

II. LINEAR MULTIUSER DETECTION

First of all a model for the system in AWGN channel is needed. The received signal for a K users system is:

$$r(t) = \sum_{k=1}^{K} s_k(t-\tau_k) + n(t) \tag{1}$$

where $n(t)$ is the single source of the channel and $s_k(t)$ is the received signal for the user k and is defined as follow:

$$s_k(t) = \sqrt{P_k}\, a_k(t) b_k(t) e^{j\theta_k} \tag{2}$$

P_k, $a_k(t)$ and $b_k(t)$ are k^{th} user's power, spreading waveform and data waveform respectively. (Both waveforms are assumed to be rectangular pulses). θ_k is the received phase of the k^{th} user relative to some reference phase.

After passing through a bank of matched filters the output of the filters can be shown to be as follows:

$$\underline{\mathbf{y}} = \mathbf{R}\mathbf{W}\underline{\mathbf{b}} + \underline{\mathbf{n}} \tag{3}$$

where $\mathbf{R}$ is a matrix with the size of $KN_b \times KN_b$ and relates

Detector	Advantages	Disadvantages
Decorrelator	• Eliminating the MAI completely • More Efficient over Conventional detector • More less complex than maximum likelihood sequence detector • Decorrelating each bit of data • Energy of signals doesn't effect the BER • Estimation of received amplitudes is not needed	• Enhancing the noise • Inverting of R matrix (which has an order of *KN* in asynchronous mode) is needed • Difficult for Real time implementation
Minimum Mean-Squared Error (MMSE)	• Taking into account the background noise • Doing a balance between noise and MAI • Generally have a better BER than decorrelator	• Estimation of users' powers are essential • The performance is depending on power of Interfering users • The near-far resistance is worse than decorrelator • Matrix inversion is needed
Polynomial Expansion (PE)	• Less Complex than decorrelator and MMSE • Can behave like decorrelator and MMSE approximately • Estimation of received amplitudes is not needed • Is applicable to both short and long codes • It's coefficients work in a large range of system parameters	• Coefficients must be calculated

Table 1. Comparison of different linear multiuser detectors

to cross-correlation of the spreading waveforms. **W** is a $KN_b \times KN_b$ diagonal matrix with square roots of the received signals' energies.

In the synchronous case and equal power signals, **R** is a $K \times K$ Auto-correlation matrix. In the conventional detector, decision is done based on $\underline{\mathbf{y}}$. In the decorrelator transformation $\mathbf{T}=\mathbf{R}^{-1}$ is used and decision is made based on $\mathbf{R}^{-1}\underline{\mathbf{y}}$:

$$\hat{\mathbf{b}} = sgn(\mathbf{R}^{-1}\underline{\mathbf{y}}) = sgn(\mathbf{W}\underline{\mathbf{b}} + \mathbf{R}^{-1}\underline{\mathbf{n}}) \tag{4}$$

For the MMSE detector a similar transformation is used:

$$\mathbf{T} = (\mathbf{R} + \sigma^2\mathbf{W}^{-2})^{-1} \tag{5}$$

The performance of MMSE approaches that of the decorrelator as $\sigma \to 0$. As σ grows larger it approaches the conventional detector. So, MMSE is a balance between decorrelator and conventional methods. However, in both cases inversion of a large matrix is needed. Ref.[3] uses a polynomial expansion (PE) of R to estimate the $\mathbf{R}^{-1}$. It uses the following transformation:

$$\mathbf{T} = \sum_{i=1}^{N} w_i \mathbf{R}^i \tag{6}$$

A comparison of different classic linear multiuser detectors in terms of their advantages and disadvantages has come in table 1.

III. ITERATIVE IMPLIMENTATION OF THE DECORRELATOR

By using the Taylor series, the expansion of $\mathbf{R}^{-1}$ can be written as [5]:

$$\alpha^{-1}\mathbf{R}^{-1} = \sum_{i=0}^{\infty} (\mathbf{I} - \alpha\mathbf{R})^i \tag{7}$$

If and only if the eigen-values of **R** satisfy the conditions:

$$\left|1 - \lambda_i(\alpha\mathbf{R})\right| < 1 \tag{8}$$

For a positive semi-definite matrix **R**, the above condition can be written as:

$$0 < \alpha < \frac{2}{\lambda_{\max}(\mathbf{R})} \tag{9}$$

Since **R** is positive semidefinite, all of its eigenvalues are positive. Based on this fact, [1] has derived an approximation for α:

$$\alpha = \frac{2}{trace(\mathbf{R})} \tag{10}$$

Because:

$$\sum_i \lambda_i = \sum_i R_{ii} = trace(\mathbf{R}) > \lambda_{\max}(\mathbf{R}) \tag{11}$$

Though it is not an accurate estimate.

IV. CONVERGENCE CONDITIONS

In a general case for the equal power and synchronous users, diagonal elements of **R** are all 1s and the non-diagonal elements are in the range of, ($-\beta$, $+\beta$), $0<\beta<1$. (β is the maximum value of the cross-correlations). **R** is symmetric and is similar to the Eq.(12):

$$\mathbf{R} = \begin{pmatrix} 1 & \beta_{1,2} & \beta_{1,3} & \cdots & \beta_{1,K} \\ \beta_{1,2} & 1 & & & \beta_{2,K} \\ \beta_{1,3} & & \ddots & & \vdots \\ \vdots & & & & \beta_{K,K-1} \\ \beta_{1,K} & & \cdots & \beta_{K,K-1} & 1 \end{pmatrix} \tag{12}$$

The condition for convergence of the Eq.(7) for every matrix is the Eq.(8). This condition must be satisfied for all of the eigen-values of matrices. Here, we examine Eq.(8) for symmetric correlation matrix **R**. For symmetric matrices, maximum and minimum of the eigen-values can be obtained by the Rayleigh-Ritz theorem [4]:

$$\lambda_{\max}(\mathbf{R}) = \max\left\{ \frac{x'\mathbf{R}x}{x'x} : x(K\times 1) real, x \neq 0 \right\} \tag{13-a}$$

$$\lambda_{\min}(\mathbf{R}) = \min\left\{ \frac{x'\mathbf{R}x}{x'x} : x(K\times 1) real, x \neq 0 \right\} \tag{13-b}$$

For our purpose, combining of Eq.(13) and Eq.(12) yields the following equation:

$$\frac{x'\mathbf{R}x}{x'x} = 1 + 2\frac{\sum_{\substack{i,j \\ i\neq j}}^{K} \beta_{i,j} x_i x_j}{\sum_{i=1}^{K} x_i^2} \tag{14}$$

The maximum and minimum of this equation gives the $\lambda_{max}(\mathbf{R})$ and $\lambda_{min}(\mathbf{R})$. When $-\beta \leq \beta_{i,j} \leq +\beta$, it is easy to show that:

$$\lambda_{\min} \geq 1 - \beta(K-1) \tag{15}$$

$$\lambda_{\max} \leq 1 + \beta(K-1) \tag{16}$$

By considering the Eq.(9, 15, and 16), following equations will be obtained:

$$1 + \beta(K-1) < 2/\alpha \tag{17}$$

$$1 - \beta(K-1) > 0 \tag{18}$$

Since **R** is a positive semi-definite matrix, Eq.(18) is not needed. (Note that Eq(17) and Eq(18) have been derived for a general case). According to Eq.(17), α can be chosen as:

$$\alpha < \frac{2}{1 + \beta(K-1)} \tag{19}$$

In the situation when the cross-correlation values are large, β takes the maximum value of (β =1) and the Eq.(19) becomes equal to Eq.(10), because:

$$trace(\mathbf{R}) = \sum_{i=1}^{K} R_{ii} = K \tag{20}$$

By using Eq(19), the performance of the system increases and the convergence rate improves.
In the synchronous cases, Eq.(19) gives a more accurate estimate for α than the one obtained from the Eq.(10) and in consequence a faster convergence can be achieved. These two methods (that use Eq(10) and Eq(19)) are simulated in synchronous environment and their performance has come in the RESULTS section.

V. GERSHGORIN ALGORITHM

In asynchronous systems most of the elements of **R** are zero while Eq.(9) considers only the maximum value of non-diagonal elements of **R** and ignores the majority of elements that are zero. This ignorance decreases the accuracy and shows its effect in the convergence rate and the performance. Another point about Eq.(9) is that it does not consider λ_{min} because the optimum value of α is:

$$\alpha = \frac{2}{\lambda_{\max} + \lambda_{\min}} \tag{21}$$

To overcome these problems, we refer to the Gershgorin theorem in linear algebra [6]. According to this theorem, any eigenvalue of a matrix is located in one of the closed discs of the complex plane centred at a_{ii} and having radius $\sum_{j,j\neq i} |a_{ij}|$. In other words,

$$|\lambda_i - a_{ii}| \leq \sum_{j,j\neq i} |a_{ij}| \tag{22}$$

By a simple calculation on the elements of **R**, two approximate values can be derived for $\lambda_{min}(\mathbf{R})$ and $\lambda_{max}(\mathbf{R})$:

$$\lambda_{max} \le \max \{a_{ii} + \sum_{j,j\neq i} |a_{ij}|\};\ i,j=1,2,...,KN_b \qquad (23)$$

$$\lambda_{min} \le \min \{a_{ii} + \sum_{j,j\neq i} |a_{ij}|\} \qquad (24)$$

Using Eq.(21), Eq.(23) and Eq.(24), a proper estimate for α can be derived.

The main benefit of Gershgorin algorithm is that it is introducing almost no complexity overhead to the detector. Its performance is also acceptable as shown in the next section.

VI. RESULTS

Some simulations have been carried out to get the performances. The first simulation concerns the synchronous users. In Fig(1) the method that uses Rayleigh-Ritz theorem for deriving the parameters (Eq(19)) is compared with the one that uses Eq(10). These two methods have been applied to a scenario containing 10 synchronous users using random codes. Both methods are implemented in 5 iterations. Their performances have been compared to the exact decorrelator, as well. It is quit obvious that the modification follows the exact decorrelator with a good degree of approximation.

Of the most interest is the performance for the asynchronous users in fading environments. Fig (2) simulates two users in fading environment having partial cross-correlations of 0.3 and 0.35 for their codes. Three methods discussed in previous section are simulated and compared with the conventional detector and exact decorrelator. These methods are implemented in 5 iterations. As can be seen, the method using Eq(10) does not have a good performance and almost performs like the conventional detector. The method using Eq(19) is not performing well, either. However, the third method that uses the Gershgorin algorithm (Eq(23, 24)) is performing appropriately and copes well in the asynchronous environments.

The iterative linear detector achieves its main performance in the early stages of operation. As shown in Fig(3), 3 stages of operation gets the main performance. This simulation is carried out for 10 asynchronous users in a 2 path fading environment having spreading codes of length 16.

The forth simulation is comparing the performance of the Gershgorin algorithm with the partial parallel interference cancellation (PPIC) method. PPIC is a well-known method with a good performance and has been of the interest of research community. Fig(4) compares the performance of an iterative linear multiuser detector using the Gershgorin algorithm to a multistage PPIC detector in 3 stages. The coefficients of the PPIC have been extracted by try and error. The scenario consists of 10 asynchronous users having random spreading codes of length 16. The channel is Vehicular A. The performance of the algorithm is slightly better than PPIC and the main benefit of the method is that

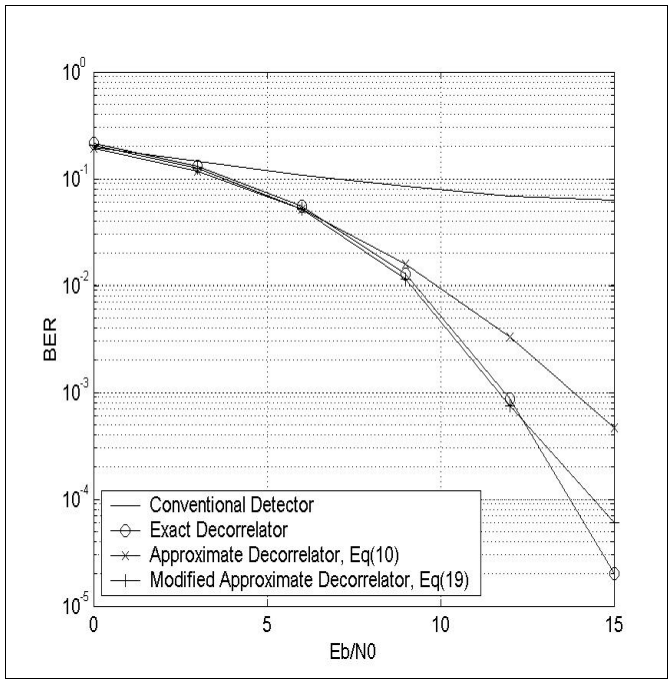


Figure 1. Decorrelator performance with 10 synchronous users, random codes, in AWGN channel and 5 stages with two conditions: simple condition Eq.(10) and modified condition Eq.(19)

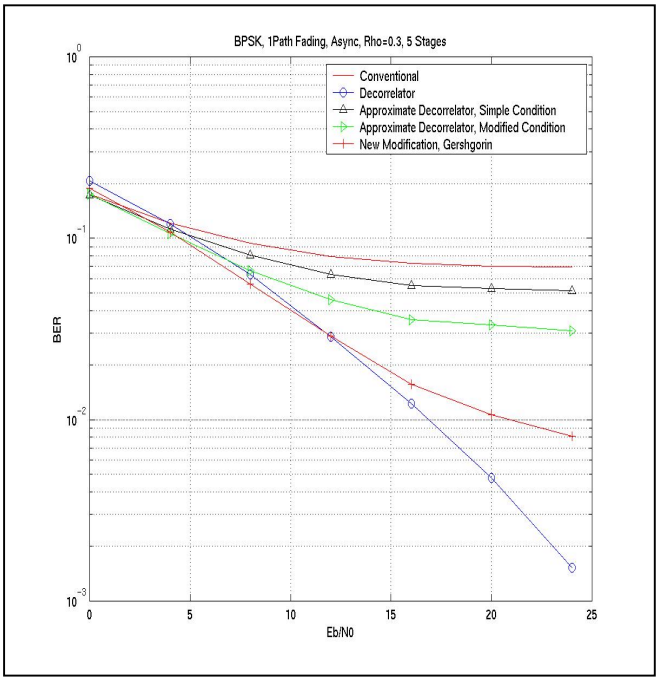


Figure 2. Decorrelator with 2 asynchronous users, Fading Channel, Crosscorrelations= 0.3, 0.35, and 5 stages with three conditions: simple condition Eq.(10), modified condition Eq.(19), and a condition based on Gershgorin theorem Eq.(21, 23, 24).

the coefficients are calculated automatically without encountering a high degree of complexity.

VII. CONCLUSIONS

In this paper a simplified method for computing the inverse of the correlation matrix has been examined and two modified versions for improving the convergence rate have been derived. The first modification was achieved by solving the equations of the convergence formulas in the synchronous environments using the Rayleigh-Ritz theorem. The second modification was even more accurate and was based on the Gershgorin theorem. The latter one increases the convergence rate of Taylor series in the asynchronous scenarios and it outperforms previous work in the literature. The performance of the iterative linear multiuser detector, that uses the Gershgorin theorem, is also compared with partial parallel interference cancellation method and performs almost the same without any need to calculate the parameters with try and error nor encountering a high degree of complexity.

REFERENCES

[1] Lei Z.D., Lim T.J., "Simplified Polynomial-Expansion linear Detectors for DS-CDMA Systems", IEE Electronic Letters, Vol. 34, No. 16, pp.1561-1563, Aug. 1998

[2] Verdu S., "Multi-user Detection", Cambridge university press, 1998

[3] Moshavi S., Kanterakis E.G., and Schilling D.L., "Multi-stage linear receivers for DS-CDMA systems", Int. J. Wirel. Inf. Newtw., 1996, 3, (1), pp.1-17

[4] M. Mozaffaripour, R. Tafazolli, "Fast Linear Multi-user Detector for DS-CDMA Systems", Capacity and range enhancement techniques for the third generation mobile communication and beyond, London, February 2000

[5] Helmut Lutkepohl, "Handbook of Matrices", John Wiley & Sons Ltd, 1996

[6] Saad Y., "Iterative methods for Sparse Linear Systems", Second edition, 2000

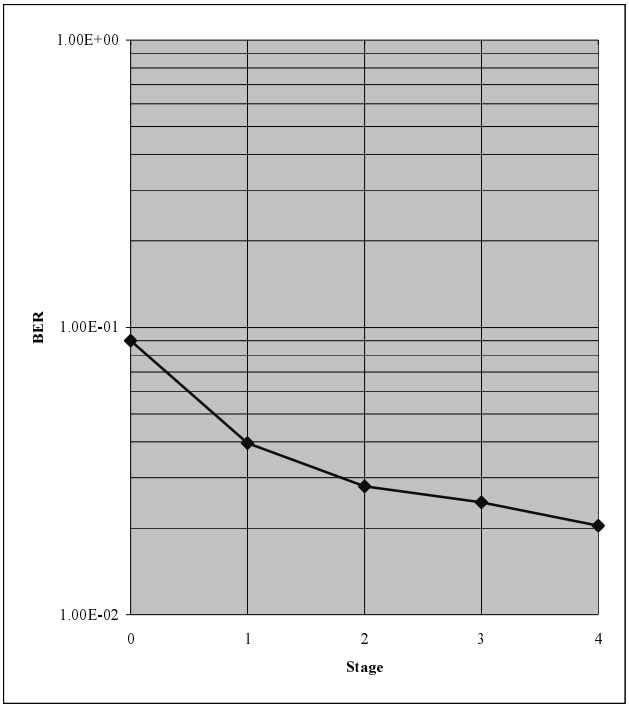


Figure 3. Multistage Effect on the performance of the iterative linear multiuser detector using Gershgorin algorithm, 10 Asynchronous Users, Random spreading codes of length 16, Vehicular A fading channel, Eb/No=14 dB

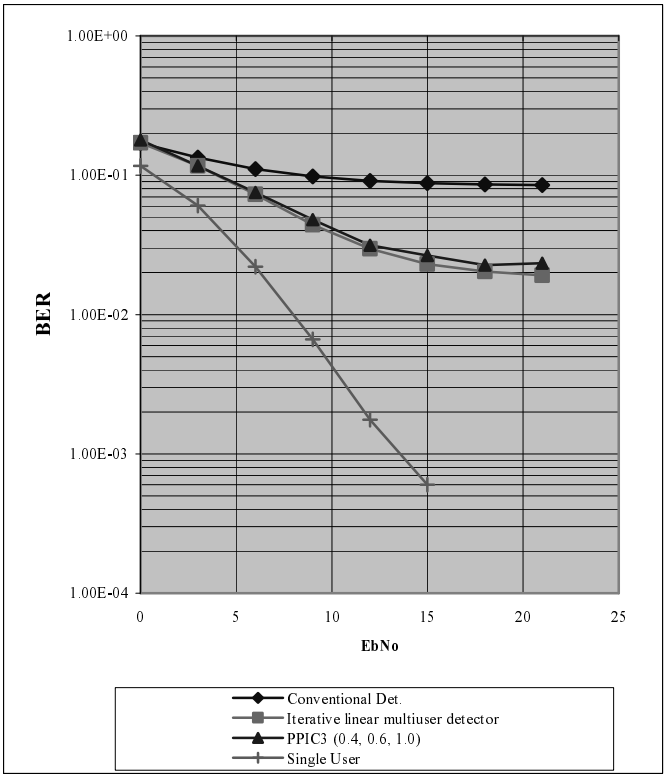


Figure 4. Performance of the iterative linear multiuser detector in compare with the PPIC detector operating in 3 stages, 10 asynchronous users, Random spreading codes of length 16, Vehicular A fading channel, Eb/No=14 dB

LOCATION ESTIMATION IN CELLULAR NETWORKS USING NEURAL NETWORKS

Javed Muhammad, Amir Hussain, W. M. Ahmed*

Dept. of Computing Science & Mathematics
University of Stirling, FK9 4LA, Scotland, UK
Corresponding Author's Email: jmu@cs.stir.ac.uk

*Dept. of Electrical Engineering, Purdue University, USA

Abstract

Location estimation finds its applications in many important decisions in cellular networks. Hand offs, cellular fraud detection and location sensitive billing are some of the examples. Many different techniques are currently in use. This paper first gives an overview of conventional location estimation techniques and applications, and a new signal-strength based neural network technique is then presented. A mobile architecture based on a simulated rural environment is used to compare the generalization performance of two types of neural networks, namely the feed forward Multi-Layered Perceptron (MLP) and the Elman Recurrent Network.

1. Introduction

Location Estimation is the process of localizing an object on the basis of some parameter. This parameter can be proximity to a detector, or some other parameter like radiated energy. The latter parameter is the one of interest in our case. In the particular context of cellular systems, this translates to the localization of the transmitter or the receiver.

Proper location estimation is very important in making many crucial decisions in cellular networks. Handoff management is one such example. When a mobile station enters from the region of service of one BS to another, a handoff is to be made. The initiation of the handoff process depends on the location of the mobile. A delay in the initiation of handoff will result in very low signal strength or in the adverse case, a call drop. Applications like handoff management don't require very accurate location estimates; all that is required is to determine which cell the mobile is in. But there are applications that ask for a very accurate estimate.

2. General architecture of a Cellular System

The cellular communication system consists of the following four major components that work together to provide mobile service to subscribers [1].

- Public Switched Telephone Network (PSTN)
- Mobile Switching Center (MSC)
- Base Stations (BS)
- Mobile Subscriber Unit (MSU)

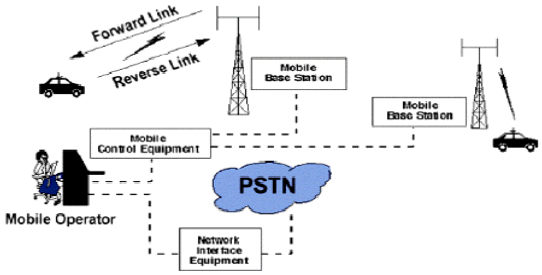


Figure 1, General architecture of a Cellular System

PSTN: The PSTN is made up of local networks, the exchange area networks, and the long haul network that interconnect telephones and other communication devices on a worldwide basis.

Mobile Telephone Switching Office (MTSO): MTSO is the central office for mobile switching. It houses the mobile switching center (MSC), field monitoring and relay stations for switching calls from cell sites to wire line central offices (PSTN). In analogue cellular networks, the MSC controls the system operation. The MSC controls calls, tracks billing information and locates cellular subscribers.

The Cell Site: The term cell site is used to refer to the physical location of radio equipment that provides coverage within a cell. The hardware located at cell site includes power sources, interface equipment, radio frequency transmitters and receivers, and antenna systems.

Mobile Subscriber Units (MSUs): The MSU consists of a control unit and a transceiver that transmits and receives radio transmissions to and from a cell site.

3. Review of Location Determination Technologies (LDT)

It is an obvious fact that locating a mobile user is much more complicated than locating the user of a fixed line. Several mobile device-positioning systems have been developed or are currently under development with different scientific approaches, the predominant of which are reviewed in this section.

Global Positioning System (GPS) is the standard location determination technology at present. A new range of LDTs are now being introduced that exploit the network of radio stations of the operators' infrastructure to locate a handset with increasing accuracy. Most network based LDTs are enhancements of the existing location capabilities of wireless networks. In fact any wireless communication is based on the ability to locate and track the position of mobile phones so that communication can be established between handsets that change position. At present LDTs fall in two main classes [2]:

3.1 Handset-based solutions are based on an active participation of the mobile terminal to the calculation of the positioning. These techniques generally require the modification of the handset as well as adding software/hardware to the cellular network.

Few handset-based solutions are available:

- Enhanced Observed Time Difference (E-OTD)
- GPS
- A-GPS (Assisted-GPS)
- Blue tooth

Handset based solutions include traditional GPS and E-OTD (Enhanced Observed Time Difference). While GPS relies on the signals of a network of satellites. E-OTD uses data received from the base stations in an operators' network to estimate the position of the handset, E-OTD requires new handset capabilities and handset software. The position calculations can be done either in the handset or in the network. The accuracy is in the range of 100 meters [2] and is far less dependent from atmospheric conditions than GPS. The main drawback of terminal based solutions is that they require new handsets.

3.2 Network-based solutions rely on positioning capabilities, which are intrinsic to the network and enable to locate an unmodified cellular phone. They require, however, the installation of specific software/hardware tools in the cellular network.

Few network-based solution are available:

- Cell ID
- Timing Advance (TA)
- Time Of Arrival (TOA)
- Angle Of Arrival (AOA)

Each technology described above has its advantages and disadvantages and when we come to evaluate a location technology the following parameters should be examined:

- Accuracy: in Rural, Urban /Dense Urban and indoor environments
- Availability: in Rural, Urban /Dense Urban and indoor environments
- Cost of implementation (both the cost impacts on handsets and/or networks)
- Technology availability
- Time to position fix

The above parameters for the positioning technologies are summarized in [7]. Some technologies have a good accuracy in one specific area and poor accuracy in other areas. On the other hand, some technologies have high availability in specific area and poor availability in other areas. The cost of implementation also varies for both the handsets and the network based solutions.

3.3 Advantages and disadvantages of handset vs. network based LDTs

Some strengths of the handset-based solutions are that it's not continuously tracked, which can also be seen as a problem to some people, and there is some privacy to calls. When the GPS chipset is added caller ID is blocked. Also, the GPS network is already active, so connecting is easy. The accuracy of this solution is better than that of the network-based solution. The accuracy for 67% of all calls is in the range of 50 meters, while at 95% it ranges to 150 meters. Some of the handset-based solutions weaknesses are that it is very expensive to purchase a GPS chipset. Also, the GPS unit depletes cell phone batteries quickly. The handset-based is not continuously tracked, as said before, it can be seen as either good or bad.

Furthermore, some strengths of the network-based solutions are that the mobile user is continuously tracked, and there is no attachment needed for the handset, which makes this solution inexpensive. There is simultaneous coverage for all subscribers, and upgrading is easy on the network. Some of the weaknesses of the network-based solutions are there is

no privacy on calls, and it's again continuously tracked. But it can only be used in certain areas. There is a high cost of deploying TDOA (time difference of arrival). The accuracy for 67% of all calls is in the range of 100 meters, while at 95% of all calls it increases to 300 meters. The accuracy is not as precise as the handset-based solution.

The E-OTD solution is best suited for GSM based networks. The E-OTD technique is already part of the GSM standard, utilizing many of the built-in standard timing measurements that the handsets already perform. A relatively simple software modification is needed in the handset to accommodate the reporting of the timing differences back to the MSC to assist in the location determination process. The TDOA and AOA methods needed to existing handsets. This means that the existing subscriber base of legacy handsets will benefit from location determination technology with no direct impact to the customer. The A-GPS method of location determination would involve the subscriber base to acquire a new GPS-equipped handset. This method would not allow legacy handsets to utilize the location determination system.

3.4 Applications of Location Technology

A brief description of some location estimation applications is presented below [3],

- Identification and exposure of people involved in cellular fraud.
- Human drivers/robot vehicles can be assisted in unknown terrains.
- Police dept in locating criminals by intercepting their telephone calls.
- Intelligent transport systems and fleet management.
- Location of stolen vehicles.

4. New signal-strength based Neural Network techniques

BS assisted or network based signal strength location estimation techniques are investigated in this work using neural networks (NN), within a simulated rural environment model. In this technique, the signal attenuation is used to estimate the distance traveled by electromagnetic waves and an estimate of the location is made. A propagation model for the signals is used. The transmitted power is known and the received power is measured. The propagation model is then used to estimate the length of the radio path. This approach is analyzed by Song in [4] who also demonstrates that multipath propagation and shadowing effects are the main sources of error in signal strength based techniques, which need to be overcome. This is currently an important area of research.

4.1 Location Estimation Using NN

This technique is akin to the use of intelligence in the cellular system. The greatest benefit is the one time training required, as the neural network is simply trained on the signal strength measurements. Collection of this data is the most laborious part of it but the rewards can be great. This one time effort can give location estimates for years until the terrain changes considerably and another training trial is required.

The general structure of a multi-layered perceptron (MLP), also known as the back propagation network, is illustrated in Figure 2, which can comprise one or more hidden layers [5].

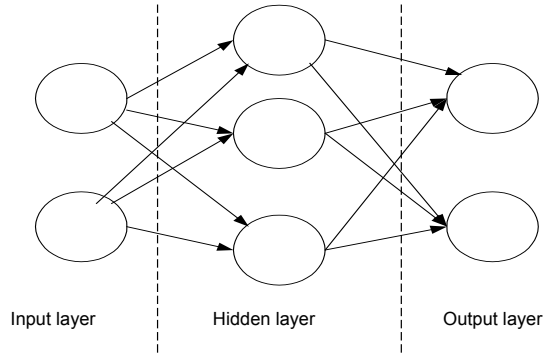


Figure2 General architecture of back propagation neural network

4.2 Mobile Architecture

The mobile architecture used for the simulations is discussed here. For the sake of simplicity, a square cell of dimensions 3km X 4km is assumed. Four BSs are used for measuring signal strengths. The coverage area is divided into grids of dimensions 0.2km X 0.2km. The idea is to place the MS in each of these grids and transmit the signal. All the four BS measure the signal

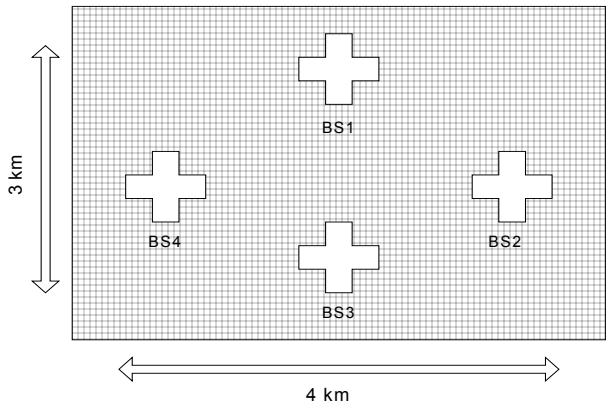


Figure 3 The square cell used for the simulation of neural network assisted location estimation

strengths for each position of the MS [3]. These measurements provide the data set for the training of the neural net.
Free space propagation equation is given below [6],

$$Pr=PtGtGr\lambda^2/(4\Pi)^2d^2L$$

Where P_r (d) is the received signal strength as a function of distance from the transmitting unit, P_t is the transmitted power, G's are the antenna gains, λ is the signal wavelength and L is the system loss factor. For the sake of simplicity it is assumed that

$$P_tG_tG_r\lambda^2/d^2L = (4\pi)^2$$

So that the free space path loss model is reduced to

$$P_r\,(d) = 1/d^2$$

With this equation signal strengths at four BS for all the positions of the MS are calculated and the neural net is trained on this data set. The origin of coordinates is taken at the left bottom corner and all measurements are taken relative to it. The coordinates of BSs are fixed and the signal strength measurement at all BSs are generated relative to the signal transmitted at the intersection of these grids.

5. Simulation Results

5.1 Using Multi-Layered Perceptron (MLP): For the situation described in figure 3, the training set consisted of 336 measurements. A two-hidden layered (4-16-16-2) MLP comprising 4 inputs, 2 hidden layers of 16 nodes each, and 2 outputs, was trained using the Levenberg-Marquardt back propagation algorithm, and the error was reduced to 0.00001 after 252 epochs, with the result that the net maps any measurement of the training set perfectly to the location of MS for that set. For test points other than the training set but within the same coverage area; the test data was generated by dividing the coverage area into grids of dimensions 0.1km X 0.1km rather than the 0.2km x 0.2km grids used to generate the training data. The test results for the MLP are given in Figure 4, for which the mobile was assumed to be at 1271 different points on the test grid. For each of these points the signal strength received at all the four BSs was calculated and was fed to the MLP which mapped these to the estimated locations. Figure 4 gives the actual and estimated X and Y coordinates, with the error in estimates (shown in Figure 5) resulting from the fact that the MLP was trained on just 363 readings. The estimation can be improved by training the net on a larger set of readings (using a smaller grid).

Figure 4

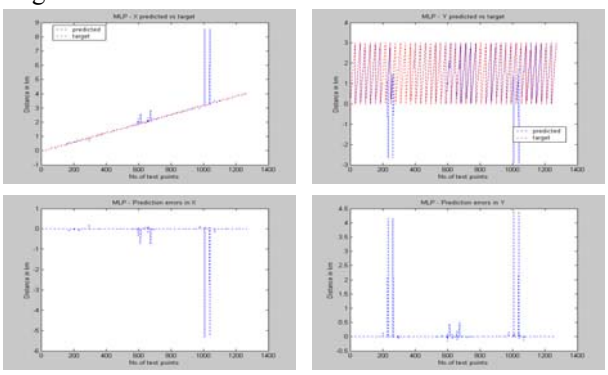


Figure 5

5.2 Using Elman Recurrent Network: Figure 6 shows the results obtained using the Elman back propagation network. The training set for this network consisted of only 154 training points generated using a comparatively larger grid-size of 0.3km x 0.3km. A two-hidden layered (4-16-24-2) Elman network comprising 4 inputs, 2 hidden layers of 16 and 24 neurons respectively, and 2 outputs, was trained using the gradient descent with momentum & adaptive learning rate back propagation algorithm, and the error was reduced to 0.00037 after 44376 epochs. For test points a grid size of 0.1km x 0.1km was used and the test results together with prediction errors are shown in Figures 6 and 7 below, from which it can be seen that the ELMAN recurrent network produces similar generalization (test) performance inspite of being trained on a much smaller (less than half) training set size but at the expense of a significantly increased training time (epochs).

Figure 6

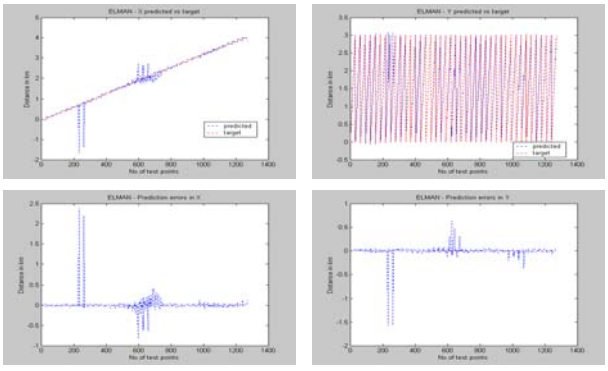


Figure 7

6. Discussion & Conclusions

The motivation behind application of neural networks to solve the location estimation problem is that the neural network technique is adept to the use of intelligence in the cellular system. The greatest benefit is the one time training. In practice however, field collection of the signal strength data is the most laborious part (and downside of the neural network approach in general) but this one-time effort can give location estimates for years until the terrain changes considerably and another training trial is required. Also, the inherent nature of the location estimation problem makes neural nets selection a wise choice for tackling this problem. Modeling the propagation of radio waves by mathematical models is quite complex involving numerous interacting variables. Multipath, diffraction and non line of sight (NLOS) cause problems. Also weather conditions affect the radio wave propagation. These are the types of problems neural networks are known to be well suited for [14], which are being investigated in this research for the location estimation problem.

New preliminary results reported in this paper extend those originally reported by Wamiq and Hussain et al. [1], which have shown that MLP and recurrent Elman based neural networks can be effectively trained on signal strength measurements obtained using a realistic simulation model for a rural environment.

For future work, the performance of the MLP and Elman networks will be compared with other types of neural networks, and the novel application of these techniques to urban areas using a modified and more realistic signal strength model will also be investigated.

7. References

[1] Cellular Communications, Web ProForum Tutorials, http://www.iec.org

[2] Location-based Services, Geo Informatics, April; 2001, http://www.geoinformatics.com.

[3] Wamiq M.Ahmed, Amir Hussain & Syed I. Shah, "Location Estimation in Cellular Networks using neural networks", Proc. International (NAISO) Symposium on Info. Sys. Intelligence. (ISI'2001), Dubai, March 2001.

[4] Han-Lee Song, "Automatic Vehicle Location in Cellular Communications Systems", IEEE Transactions on Vehicular Technology, Vol 43, No.4, November 1994.

[5] Haykin, S. (1994) *Neural Networks: A Comprehensive foundation.* Upper Saddle River, NJ: Prentice Hall.

[6] Theodore S. Rappaport, "Wireless Communications Principles & Practice", Prentice Hall, 1996.

[7] J. Muhammad, *Location Estimation in Cellular Networks: A critical review of conventional techniques*, Technical Report, Computing Science & Maths, Stirling University, Dec 2002.

DECISION DEPTH OF SLIDING WINDOW SOFT-INPUT SOFT-OUTPUT ALGORITHMS

R. Hoshyar[1], S. H. Jamali[2], A. R. S. Bahai[3], and R. Tafazolli[1]
[1]CCSR, University of Surrey, UK, r.hoshyar@surrey.ac.uk, r.tafazolli@surrey.ac.uk,
[2]University of Tehran, IRAN, hjamali@chamran.ut.ac.ir,
[3]National Semiconductor, and Stanford University, US, bahai@stanford.edu,

ABSTRACT: *The glory of the concatenation systems, such as turbo codes, stands not only on the efficient use of the concatenation to improve the code structure, but also on their sub-optimum iterative decoding. The main ingredient in iterative decoding is the Soft-Input Soft-Output (SISO) algorithm. The optimum SISO algorithm is realized by MAP (Maximum A Posteriori), or its logarithmic version LogMAP, algorithm. MAP suffers from memory usage in the case of large code length. Sliding window (SW) versions of MAP, and also SOVA (Soft Output Viterbi Algorithm) are sub-optimum ones with less memory requirement and close to optimum performance. The decision depth of these algorithms should be adjusted to achieve the appropriate performance-complexity trade-off. Here simulation results will be provided on the effect of decision depth on the performance of both SW version of MAP, and SOVA.*

I. INTRODUCTION

Turbo coding presented first in [1] has attracted the most for its near capacity performance. Turbo code is the parallel concatenation of two or more convolutional codes. Concatenation of simple constituent codes through appropriate interleaving provides a good opportunity to achieve an overall coding system with an excellent weight structure. Such a coding system is appreciated if only a realizable method is presented for its decoding. Fortunately this was done for turbo codes by presenting an iterative decoding approach. The main ingredient in iterative decoding is the Soft Input Soft Output (SISO) algorithm. This algorithm gets the necessary reliability information and by imposing the constituent code constraint generates new reliability information. Iterative method tries to jointly impose all constituent codes constraints by exchanging the reliability information among the corresponding SISO modules.

The main complexity of the iterative decoding is due to SISO algorithms, and there have been noticeable efforts to present and implement low complexity ones. SISO algorithms belong to one of the two families of MAP [2] and SOVA ([3], and [4]) algorithms. MAP is the optimum SISO and computes a posteriori probabilities of the constituent symbols. Its logarithmic version LogMAP also is optimum and is numerically more stable than MAP. MAP family has sub-optimum sliding window (SW) versions that avoid large storage with some computation overhead. SOVA algorithms are modification of the VA. The main idea in these algorithms, is the sub-optimum computation of the constituent symbols' a posteriori probabilities by the use of the side information existing in the classic VA.

II. SOFT INPUT SOFT OUTPUT DECODING

Based on the received sequence $\underline{Y}$, and coding constraint, that the transmitted sequence $\underline{C}$ must belong to the set of codewords *C*, a posteriori probability of the binary symbol $u_i^{(j)}$, the *j*-th bit of the *i*-th trellis section, is computed as follows:

$$l\left(u_i^{(j)};O\right)=ln\left\{\frac{P\left(u_i^{(j)}=+1|\underline{Y},C\right)}{P\left(u_i^{(j)}=-1|\underline{Y},C\right)}\right\} \quad (1)$$

Partitioning the codewords set *C* to the subsets $C_{i,j}^{+1}$ and $C_{i,j}^{-1}$, consisting of all the codewords with, $u_i^{(j)}=+1$ and $u_i^{(j)}=-1$, respectively, the soft value $l\left(u_i^{(j)};O\right)$ can be computed as follows:

$$l\left(u_i^{(j)};O\right)=\ln\left\{\frac{\sum_{\underline{C}\in C_{i,j}^{+1}}P(\underline{C},\underline{Y})}{\sum_{\underline{C}\in C_{i,j}^{-1}}P(\underline{C},\underline{Y})}\right\} \quad (2)$$

Using the code trellis and further partitioning of the sets $C_{i,j}^{+1}$ and $C_{i,j}^{-1}$, to the subsets that their codewords passing through different states, MAP algorithm computes above soft output value through the following forward and backward recursions:

$$l\left(u_i^{(j)};O\right)=ln\left\{\sum_{(s',s):u_i^{(j)}=+1}p(s',s,\underline{Y})\Big/\sum_{(s',s):u_i^{(j)}=-1}p(s',s,\underline{Y})\right\},$$

where

$$p(s',s,\underline{Y})=\alpha_i(s')\cdot\gamma_i(s',s)\cdot\beta_{i+1}(s), \quad (3)$$
$$\alpha_{i+1}(s)=\sum_{s'}\alpha_i(s')\cdot\gamma_i(s',s),$$
$$\beta_i(s')=\sum_{s}\gamma_i(s',s)\cdot\beta_{i+1}(s),$$

where $\alpha_i(s)$ and $\beta_i(s)$, are accumulated forward and backward probabilities of state *s* at time index *i*. $\gamma_i(s',s)$ is the probability of branches $(s',s)_i$ (the branches that make transition from state s' to state *s* at

time index *i*). Initialisation of above recursions should be done based on the initialisation and termination of the trellis structure. Applying logarithm to above recursions, sum-product MAP algorithm will change to max*-sum LogMAP algorithm, which is numerically stable. where the operator max* is defined as $\max_i^*(a_i) \stackrel{\Delta}{=} ln\sum_i e^{a_i}$. Max* can be approximated by max operator. Replacing max* with max operator in the LogMAP algorithm, its sub-optimum version SubMAP will be resulted, that performs close to LogMAP at high signal to noise ratios (HSNR).

MAP suffers from the large storage requirement when working on long codes. As one of the forward or backward parameters must be stored for all the code length.

III. Sliding Window SISO

Trellis constraint along with the noisy behaviour of the channel induces correlation among the received neighbouring symbols $\underline{Y}$. Span of this correlation depends on the trellis width, its number of states or code memory, and the noisy condition of the channel. A symbol is mainly correlated to its neighbouring symbols within a span, and in the inference of its reliability value the other symbols may safely be ignored. Consequently equation (1) can be approximated by truncating the codewords up to length $i+\Gamma$, where Γ is the appropriately selected decision depth:

$$l\left(u_i^{(j)};O\right) \approx ln\left\{\frac{P\left(C_{i,j}^{+1}(\Gamma)\middle|\underline{Y}_0^{i+\Gamma}\right)}{P\left(C_{i,j}^{-1}(\Gamma)\middle|\underline{Y}_0^{i+\Gamma}\right)}\right\} . \quad (4)$$

$C_{i,j}^{\pm 1}(\Gamma)$ is the set of partial codewords up to $i+\Gamma$ trellis section with $u_i^{(j)} = \pm 1$, and $\underline{Y}_j^l$ is the portion of the received sequence from *j*-th trellis section up to *l*-th trellis section. Indeed the above equation is identical to (1) for a non-terminated trellis of length $i+\Gamma$; therefore, we can use similar recursions of the original MAP algorithm with the exception that for the soft value computation of a symbol belonging to the $i-th$ trellis section, its corresponding backward recursion starts from time index $i+\Gamma$:

$$\alpha_{l+1}(s) = \sum_{s'} \alpha_l(s') \cdot \gamma_l(s',s) , \qquad l = 0,1,\cdots \quad (5)$$

$$\beta_l^{(i)}(s') = \sum_s \gamma_l(s',s) \cdot \beta_{l+1}^{(i)}(s),$$

$$l = i+\Gamma, i+\Gamma-1, \cdots, i+1,$$

where $\beta_l^{(i)}(s)$, $l = i+\Gamma, i+\Gamma-1, \cdots, i+1$, are the backward parameters used in the soft value computation of the $i-th$ trellis section. Similar to the original MAP, initialisation of the forward recursion depends on the trellis initialisation, but for the backward recursion the trellis has not yet been terminated and it can be uniformly initialised or use forward parameters at time index $i+\Gamma+1$ for its initialisation [5]. Similarly sliding window (SW) versions of LogMAP, and SubMAP can be readily developed.

IV. Soft Output Viterbi Algorithm

SOVA sub-optimally computes equation (1) using the side information that exists in the classic VA. Approximating max* operator with max, equation (4) will take the following form:

$$l\left(u_i^{(j)};O\right) \approx \max_{\underline{C} \in C_{i,j}^{+1}(\Gamma)} \left\{ln\ P\left(\underline{C}\middle|\underline{Y}_0^{i+\Gamma}\right)\right\} - \max_{\underline{C} \in C_{i,j}^{-1}(\Gamma)} \left\{ln\ P\left(\underline{C}\middle|\underline{Y}_0^{i+\Gamma}\right)\right\}. \quad (6)$$

Denoting $\hat{u}_{i,MAP}^{(j)} = \hat{u}$ as the corresponding bit of the codeword with maximum a posteriori probability, $\hat{\underline{C}}_{MAP}$, above equation can be expressed as follows:

$$l\left(u_i^{(j)};O\right) \approx \hat{u} \cdot \min_{\underline{C} \in C_{i,j}^{-\hat{u}}(\Gamma)} \left\{\ln P\left(\hat{\underline{C}}_{MAP}\middle|\underline{Y}\right) - \ln P\left(\underline{C}\middle|\underline{Y}\right)\right\}. \quad (7)$$

$C_{i,j}^{-\hat{u}}(\Gamma)$, is the set of partial codewords up to time index $i+\Gamma+1$ that their corresponding bit is $u_i^{(j)} = -\hat{u}$, which is not equal to the corresponding bit of MAP codeword $\hat{u}_{i,MAP}^{(j)} = \hat{u}$. The expression to be minimized in equation (7) is the metric difference of the MAP codeword $\hat{\underline{C}}_{MAP}$ and some member of $C_{i,j}^{-\hat{u}}(\Gamma)$. If this member, which is also a path of the trellis diagram, merges to MAP path at some state *s* and is equal after that, it has largest metric among all of the codewords of $C_{i,j}^{-\hat{u}}(\Gamma)$ passing through the same state *s*. Thus if the search in minimization of (7) is only confined to a subset of $C_{i,j}^{-\hat{u}}(\Gamma)$, that its members cross MAP path in one or more states, an approximation of (7) will be resulted, that partial metric difference can be safely used instead of the total metric difference. During the classic VA some members of this subset may compete with MAP codeword in the ACS operation of VA. Further confining of the minimization to only such codewords another approximation will be yielded:

$$l\left(u_i^{(j)};O\right) \approx \hat{u} \cdot \min_{\underline{C} \in C_{i,j}^{-\hat{u}}(SOVA,\Gamma)} \left\{\Delta_i^{(j)}\right\}. \quad (8)$$

where $C_{i,j}^{-\hat{u}}(SOVA,\Gamma)$ is a subset of $C_{i,j}^{-\hat{u}}(\Gamma)$, that its members have competition with MAP path in classic VA, and $\Delta_i^{(j)}$ is the partial metric difference between the MAP path and a competing path that its $u_i^{(j)}$ bit differs with that of MAP path. Thus by taking the following procedure, VA will be able to produce soft value of decoded bits. During each ACS operation the partial metric difference between two competing paths is computed. Then the corresponding constituent symbols of the two paths are compared to each other. If

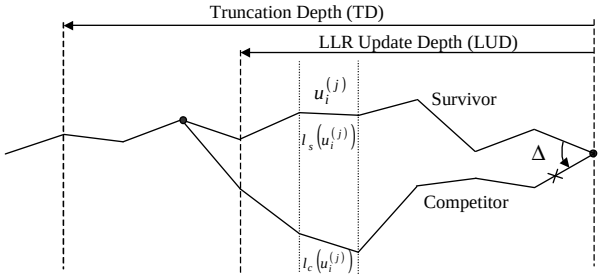


Figure 1. Back search processing in SOVA

they are unequal and the observed metric difference is lower than the previously registered metric difference in the survivor path, the previously registered value will be substituted by the observed metric difference. Therefore, by addition of the above back search processing to the VA, this algorithm will be equipped with soft value computation capability. The depth used in the back search processing, called LLR Update Depth (LUD), is very important in both accuracy and complexity of SOVA. This depth can be safely set lower than the Truncation Depth (TD), which is the Decision Depth for hard value computation of the VA. Figure 1 illustrates the back search processing of SOVA, and its TD and LUD. $l_s\left(u_i^{(j)}\right)$, and $l_c\left(u_i^{(j)}\right)$ are the tentative LLRs of the survivor and competitor paths corresponding to bit $u_i^{(j)}$. New LLR value for the survivor path is computed as follows:

$$\left|l(u_i^{(j)})\right| = \begin{cases} \min\left\{\left|l_s(u_i^{(j)})\right|, |\Delta|\right\}, & \operatorname{sgn}\left(l_s(u_i^{(j)})\right) \neq \operatorname{sgn}\left(l_c(u_i^{(j)})\right), \\ \left|l_s(u_i^{(j)})\right|, & \operatorname{sgn}\left(l_s(u_i^{(j)})\right) = \operatorname{sgn}\left(l_c(u_i^{(j)})\right), \end{cases}$$
$$\operatorname{sgn}\left(l(u_i^{(j)})\right) = \operatorname{sgn}\left(l_s(u_i^{(j)})\right) \quad (9)$$

V. Decision Depth Analysis

Computation of (2) requires the reception of the entire sequence $\underline{Y}$, which besides imposing the delay problem requires a large amount of memory for large block sizes. Also in the applications such as non-terminated convolutional codes, block size is infinite, while a practical decoder must output its result within a finite delay. As discussed in section III, neighbouring symbols are the main sources of inference for soft value computation of a symbol. This led us to the approximate equations of (4), and (8).

SW versions of MAP, and SOVA do not require large storage, and their required amount of memory is independent of the code length. These algorithms provide this advantage by some extra processing. The decision depth parameter determines the complexity overhead of these algorithms. The larger the decision depth, the more the complexity overhead, and the better the performance. This parameter should be selected appropriately to achieve close to optimum performance with minimized complexity overhead. Performance-complexity optimisation of SOVA requires adjustment of both hard decision depth TD and soft update depth, LUD.

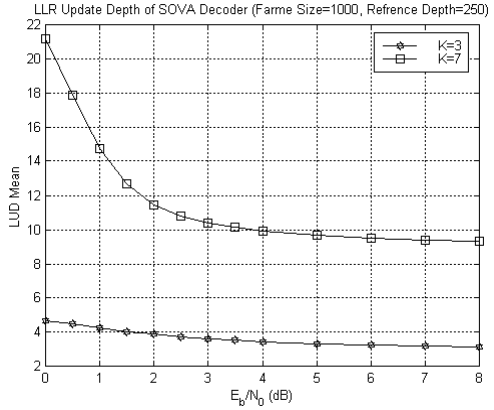
Using the theory of the product of random matrices, reference [6] has provided decision depth analysis for a specific type of a sliding window MAP algorithm. The analysis result is applicable to all SW versions of MAP. It shows that the required decision depth depends on the noisy condition of the channel, and trellis width of the code (number of states). The noisier the channel, and the larger the trellis width, the larger decision depth will be required.

The decision depth analysis presented in [6] is also applicable to SOVA family SISO decoders. Indeed SOVA family uses an approximation of equation (6), which SW-SubMAP is based upon. LLR Update Depth (LUD) simulation is carried out for SOVA decoding for two half rate convolutional codes with constraint length 3, and 7, and octal generators (7,5), and (554,744), respectively. A SOVA decoder with large LUD=100 is used as a reference, and another SOVA is used as a test decoder. In this decoder the ML path is exactly found using classic VA, then for this path back-search LLR updating of SOVA is applied. In each LUD depth of d the LLR $l(d)$ is compared with its corresponding reference value $l(ref)$, computed by reference decoder. If they are in ε=0.001 neighbouring of each other, i.e. $|l(ref)-l(d)|<\varepsilon$, corresponding depth will be registered. Similar simulation is carried out for FSW-MAP and FSW-LogMAP decoders, where FSW (Forward Sliding Window) is a SW version presented in [6] and is more convenient for decision depth simulation. All of the results confirmed that the required decision depth and LLR update depth increases with decreasing of SNR, and increasing of code constraint length.

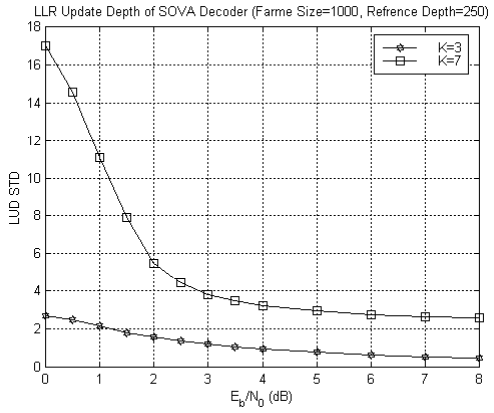
Figure 2 shows the simulation results of the mean, and Standard Deviation (STD) of the required LUD for SOVA decoder versus SNR for the two codes, with constraint length 3 and 7, respectively. Figure 3 shows the mean, and Standard Deviation (STD) of the required LUD for SOVA decoder for the two cases of Non Recursive (NR) and Recursive Systematic Convolutional (RSC) codes with constraint length 3. Also shown in this figure is simulation of the decision depth mean, and STD of the RSC code with K=3 and FSW-MAP, FSW-MAP with MC (maximum component) estimation [6], FSW-LogMAP, and FSW-SubMAP decoders. As it is expected FSW-MAP and FSW-LogMAP have the same results. Due to more accurate computation these decoders need more decision depth than their SubMAP sub-optimum version. FSW-MAP with MC estimation lies halfway between them. This is justified by noting that this decoder is equivalent to an FSW-LogMAP decoder whose soft value computation is carried out using *max* instead of max* operator.

Except a bias term, the mean plots related to SOVA and FSW-SubMAP have the same forms. This difference can be justified as follows. The structure of the code and channel SNR imply a typical set of error events; a minimal set that with a probability near to one the occurred error events belong to this set. LUD of SOVA decoder must be adopted such that in most of the cases the lengths of these typical error events are shorter

(b)

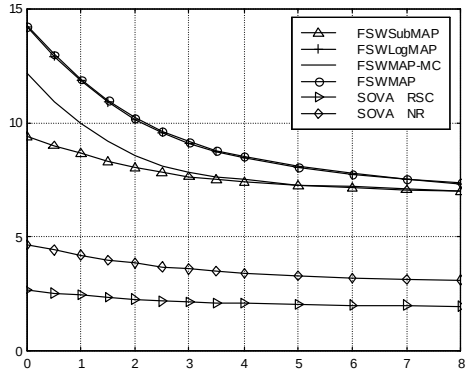


(a)

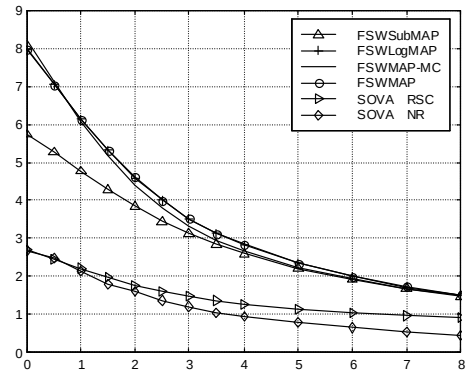


(b)

Figure 2. LUD depth simulation of SOVA decoder over AWGN channel for the two rate-1/2 convolutional codes with octal code generators (5,7), K=3 and (554,744), K=7 over AWGN channel, respectively. (a) Mean, (b) STD (standard deviation).



(a)



(b)

Figure 3. LUD and decision depth simulation of FSW-MAP, FSW-MAP with MC estimation, FSW-LogMAP, FSW-SubMAP, and SOVA decoders for a rate-1/2 RSC code with octal code generator (1,5/7), K=3 along with those of SOVA decoder for Non Recursive (NR) convolutional code (7,5) over AWGN channel: (a) Mean, (b) STD (standard deviation).

than this depth. Due to the finite length of survivor paths, truncation depth (TD), it is possible that instead of survivor path corresponding to ML path, another survivor path achieves maximum accumulated metric, while its tail has not merged to the ML path yet. We call such an event as partial error event.

Figure 4 shows partial and full error events. Although the partial error events have the potential of being an error event, but they are not necessarily an error event as they do not occur in optimum ML decoding. This is deficiency of decoder because of low TD that these errors occur. With occurrence of such errors, although the ML survivor path has possibly gathered good LLR values, the inaccurate LLRs of the winner survivor will be delivered as soft output value. The average length of partial error events is greater than the average length of error events, as they can have longer excursions through the trellis. Therefore, in SOVA family TD must be selected greater than LUD.

Let's consider FSW-SubMAP algorithm, in the case of a large enough Γ, the restricted forward state metric distributions, for the two values $u_i^{(j)} = \pm 1$ will converge to a same distribution except with an additive bias term [6]. This convergence implies that all of the possible error events with error on $u_i^{(j)}$ have merged to the ML

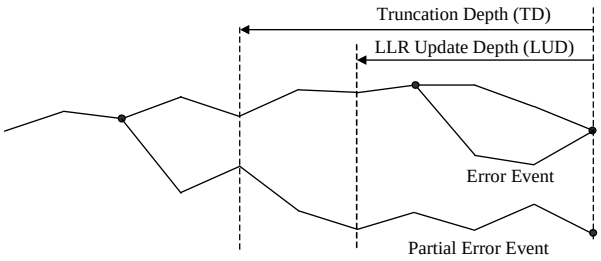


Figure 4. Partial and full error events in VA and SOVA family

path before all possible partial error events diverge from MAP path, and their effect is only a bias term on state metric distribution (It takes a little brainwork). As illustrated in Figure 5, occurrence of an error event will have an additive bias effect on metric distribution over the states if and only if all partial error events diverge from MAP path after this error event. This leads us to an interesting consequence that the necessary decision depth for sliding window versions of LogMAP algorithm must be greater than the necessary truncation depth in VA and SOVA, while both keep the same form of variation in terms of SNR and code memory.

III. Simulation Results

A rate 1/3 turbo code (Parallel Concatenated Convolutional Code) with two constituent systematic recursive (RS) convolutional codes, and interleaver size of 1024 over AWGN channel, and BPSK modulation was simulated. The constituent codes have octal representation of (1,5/7). Figure 6.a shows the 10^{th} iteration BER performance of the considered turbo code for different decision depths of SW-SubMAP decoder. As it is seen there are considerable performance loss for low values of decision depths. Decision depth of 16 is appropriate and presents the same performance as SubMAP decoder. BER simulation results of this turbo code system for the 10^{th} iteration and various LUDs of TSOVA decoder are shown in Figure 6.b. TSOVA (Threshold SOVA) is a version of SOVA that clips the output of SOVA to control its optimistic behaviour [7]. Similar to SW-SubMAP low values of LUD cause considerable performance degradation. LUD value of 8 has nearly the same performance as greater values, and is an appropriate choice. It is noticeable that the necessary value of LUD for TSOVA is much lower than the necessary decision depth for SW-SubMAP decoder. The required decision depth is almost the same for both algorithms, but the required value of the LUD of SOVA is almost half of the required decision depth for the same condition.

IV. Conclusion

Sliding window versions of MAP, and SOVA are more appropriate for iterative decoding implementation. These SISO algorithms do not impose large storage requirement, and provide close to optimum performance. These algorithms instead require more processing than original MAP algorithm. The decision depth is the most important parameter in these algorithms that provides a fine trade off between the processing load and the performance. Simulation results show that the required LUD of SOVA is almost half of the necessary decision depth of sliding window versions of MAP.

V. References

[1] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon Limit Error-Correcting Coding and Decoding, Turbo Codes," *in Proc. ICC'93 Geneva Switzerland*. May 1993, pp.1064-1070.

[2] L. R. Bahl, J. Cocke, F. Jelinek, and Raviv, "Optimal Decoding of Linear Codes for Minimizing Symbol error Rate," *IEEE Trans. Inform. Theory, Vol. IT-20*, pp. 284-287, Mar. 1974.

[3] C. Berrou, P. Adde, E. Angui, and S. Faudeil, "A Low Complexity Soft-Output Viterbi Decoder architecture," *in Proc. ICC'93 Geneva Switzerland*, pp.737-740, May 1993.

[4] J. Hagenauer, and P. Hoeher, "A Viterbi Algorithm with Soft-Decision Outputs and its Applications," *in Proc. IEEE Globecom Conf. Dallas*, TX, pp. 1680-1686, Nov. 1989.

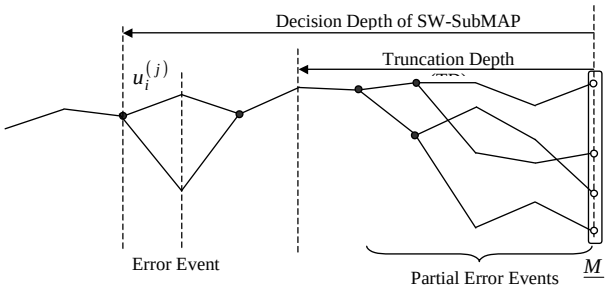


Figure 5. Relation between truncation depth of VA and decision depth of SW-LogMAP. The occurrence of error event has only an additive bias effect on metric distribution $\underline{M}$. This is only possible when all partial error events diverge from ML path after this error event.

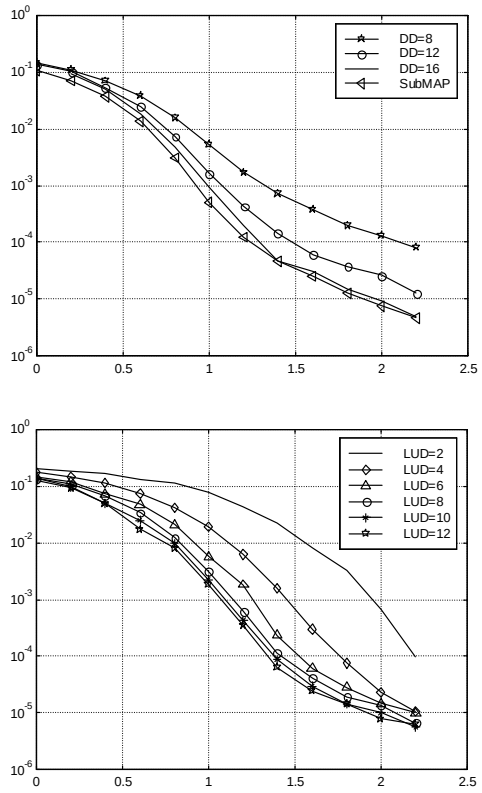


Figure 6. BER simulation of the considered turbo code for different (a) decision depths of SW-SubMAP decoder, and (b) LLR update depths (LUD) of TSOVA decoder.

[5] S. Bendetto, D. Divsalar, G. Montorsi, and Pollara, "A Soft-Input Soft-Output Maximum a Posteriori (MAP) Module to Decode Parallel and Serial Concatenated Codes," *The Telecommunications and Data Acquisition Progress Report 42-127*, Aug. 1996.

[6] X. Wang, and S. B. Wicker, "A Soft-Output Decoding Algorithm for Concatenated Systems," IEEE Trans. Inform. Theory, Vol. 42, No. 2, pp. 543-553, Mar. 1996.

[7] L. Lin, and R. S. Cheng, "Improvements In SOVA-Based Decoding For Turbo Codes," *in Proc. IEEE ICC97*, pp. 1473-1478.

Joint Estimation of Signal and Interference Power in Rayleigh Fading Channels

I. Introduction

ROBUST joint estimation of signal and interference is essential for reliable communication over fading channels in mobile wireless systems. To accomplish this objective, both receiver and transmitter require knowledge of signal and interference power for adaptation purposes. These quantities can vary substantially depending on channel condition and must therefore be estimated accurately. A typical application for signal and interference power estimation is in the transmit power control (TPC) in fixed or mobile wireless environments. TPC allocates transmit power among users to minimize interference between adjacent users while maintaining the signal quality. Another immediate application of signal and interference power estimation is in likelihood estimation in soft decoding algorithms, such as Turbo decoders.

Any imperfection in estimating signal and noise power, whether introduced through bias or large variance in the estimate, can have a dramatic and sometimes catastrophic effect on the system performance. We will concentrate solely on estimators that provide low variance estimates while maintaining the residual bias within a reasonable range.

Estimation of signal and noise power in wireless fading channels has long been a topic of research. Specific to wireless systems, most of the research has followed two distinct analysis path. In the first, a linear Baysean type estimator, such a Kalman or Wiener filter, was proposed for estimation purposes. Jiang *et al.* , for example , modelled the shadow process with a first order AR model [3]. Tanskanen *et al.* , suggest using a Wiener filter for transmit power control (TPC) mechanism in WCDMA systems [10]. The second approach which is more of a classical estimation approach, derives the estimators based on statistical or rather probabilistic characteristic of the observation space. An example of this approach can be found in [9] and [13]. Focusing solely on the Kurtosis of the observation vector, Ramesh *et al.* in [6] provided an estimator based on higher order statistics. The proposed estimator relies on the diversity reception that is based on multiple receive antenna structure in the receiver. In [5] Pauluzzi and Beaulieu have provided a comprehensive study on statistical analysis of SNR estimation methods for AWGN channels.

These methods provide very useful results. There is, however, a need for analysis techniques that provide a unified framework for analyzing signal and power estimation in Rayleigh fading environment. Our approach for solving this problem is more of the classical estimation, rather than Baysean philosophy. This decision is well justified knowing if we take into our consideration that the underlying unknown parameters, i.e. signal and interference power, are not a linear function of the observation space. Naturally, applying linear MMSE Baysean estimators would result in unwanted bias or even large variance in estimation of the unknown parameters.

We aim to provide a solution to this problem under the reference symbol assisted (pilot based) scenario as opposed to non data-aided (blind) methods for the following reasons. Blind methods would not necessarily result in unbiased estimation of the unknown parameter and also suffer from slow convergence at low SNR. Moreover, pilot-aided reference training is suitably applicable to most mobile wireless systems such as 3G WCDMA. The insertion of dedicated pilot channel (DPICH) in the forward link of the WCDMA systems allows partial knowledge of the noise power parameter at the receive ends assuming that channel condition stays fairly constant over consecutive pilot symbol interval (sufficiently slow fading and mobile movements). In this scenario, the interference variance estimation problem would be a special case of noise variance estimation problem in partially-known signals. However, the deep fades and rapid phase changes in wireless mobile channels make estimation of the interference variance quite difficult. We have tailored our analysis to address this issue for robustness purposes. In doing so, we formulate the problem as a point estimation of a multivariate Gaussian kernel density function. As a member of the exponential class of distributions, Gaussian distribution admits a sufficient statistic. Such a statistic is intimately related to the concept of maximum likelihood estimation (MLE) and minimum variance unbiased estimation (MVUE) of unknown parameters. We use this property to derive from the samples of pilot observation space, a class of estimators, namely MLE, MVUE and sub-space estimator, that provide optimum characteristics in the Cramer-Rao sense. The proposed estimators do not impose any requirement on structure of the receiver and are applicable to both diversity and non-diversity based receivers. Moreover, there is no underlying assumption to the spectral characteristic of the Rayleigh fading process.

The paper is organized as follows. In the next section, we formulate the problem and outline the Rayleigh fading channel model. In section III, we briefly discuss the proposed unified class of estimators for signal and interference power estimation. And finally we provide some conclusions and simulation results. Further statistical analysis and details of how these estimators will be applied to Wideband CDMA systems in practice will be given in the full version of the paper.

II. Problem Definition

Consider communication system that employs pilot symbol assisted symbols that operate over a fading channel. The channel consists of two components, a linear time varying filter which captures the effects of multipath fading due to multiple scatters

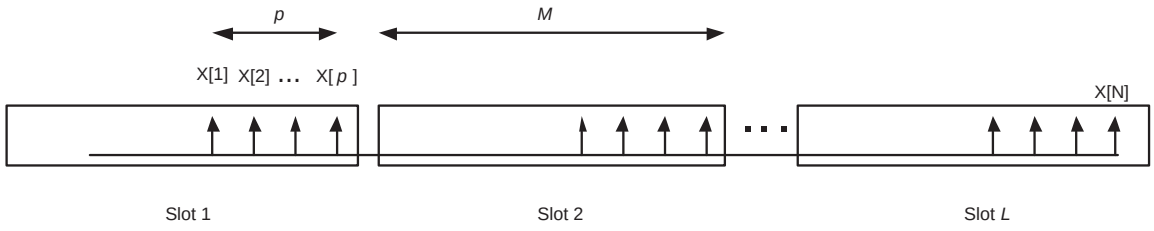


Fig. 1. Transport Formats Structure in 3GPP WCDMA

in the transmission medium, and an additive white Gaussian noise(AWGN) term representing both receiver noise and, more significantly, co-channel interference.

More specifically, the response of the channel to a known reference sequence $z[n]$ is given by

$$r[n] = \sum_k a[n;k]z[n-k] + u[n] \tag{1}$$

where $z[n]$ is the complex valued M-ary symbol sequence to be transmitted and $u[n]$ is a complex zero-mean circularly symmetric Gaussian random process.

In this paper, we focus our discussions on CDMA systems, even though a slight modification of the same arguments can be used for Single-Carrier Modulation systems as well as OFDM systems. For CDMA systems, assuming an ideal performance, the Rake receiver resolves the received signal into its multipath components. Hence, for the a-priori known transmitted Pilot symbols the output of each Rake finger after derotation by the conjugate of the known sequence, can be represented as

$$x[n] = \sqrt{\rho}\alpha[n] + w[n] \tag{2}$$

wherein $w[n]$ is also a complex zero-mean Gaussian random process with an unknown variance of σ^2. The real and imaginary parts of the complex fade coefficients $\alpha[n]$ are the unit power independent Gaussian random processes with known autocorrelation function $\varphi[n]$, a well-known example of which is the Bessel function for the Rayleigh fading model given by the Jakes model (see [8]). Here, without loss of generality we assume that the fade coefficients have unit variance.

Subsequently, we define the observation space as a sequence of the above received symbols $x[n]$, corresponding to the known pilot symbols. Note that for the special case of 3G Wideband CDMA systems, a sequence of a-priori known *dedicated Pilot symbols* are periodically time-multiplexed with the data symbols based on a specific *transmission slot format* as shown in Figure (1). In this case, the observation vector will consist of L such intervals.

Now, given N samples of observation space $x[n]$, our objective is to obtain the point estimation of the signal power $\rho$and interference variance σ^2. The sufficient statistics for estimation of the noise and signal power, σ^2 and ρ, is obtained by forming N successive samples of the received pilot symbols $x[n]$ into $\bar{x} = [x[1]\ x[2]\ , \cdots x[N]]^T$. Then, under the above assumptions the probability density function of the observation vector is a multivariate Gaussian mixture given by

$$pr(\bar{x};\sigma^2) = \frac{1}{(2\pi)^{\frac{N}{2}}|R_x|} exp(-\bar{x}^* R_x^{-1}\bar{x}) \tag{3}$$

In the above equation, the covariance matrix of the observation vector $\bar{x}$ is defined as

$$R_x = \rho R_\alpha + \sigma^2 I \tag{4}$$

where entries of the fading covariance matrix R_α are defined as $R_\alpha[m,n] = \varphi[m-n]$

Due to Hermitian property of the auto-covariance matric $R_{\bar{x}}$, there exist an eigenvalue decomposition transform such that

$$R_\alpha = U\Gamma U^* = \sum_{i=1}^{N} \gamma_i u_i . u_i^* \tag{5}$$

By definition, columns of the matrix U consist of orthonormal eigenvectors u_i that make the matrix U to be unitary. Matrix and $\Gamma = diag[\gamma_1\ \gamma_2\ \cdots \gamma_N]$ is a diagonal matrix with diagonal entries equal to real non-negative eigenvalues of matrix R_α. From our assumption that the variance of α is equal to 1, it is easy to see that

$$\sum_{i=1}^{N} \gamma_i = \text{Trace}(R_\alpha) = N \tag{6}$$

Thus, the inverse of auto-covariance matrix can be written as

$$R_x^{-1} = (\rho U \Gamma U^* + \sigma^2 I)^{-1} = \sum_{i=1}^{N} \frac{u_i.u_i^*}{\rho\gamma_i + \sigma^2} \tag{7}$$

Substituting equation (7) in (3) and using the identity $|R_\alpha| = \Pi_{i=1}^{N}\gamma_i$, the likelihood function is written as

$$\Lambda(\bar{x}; \sigma^2 | f_D) = \ln pr(\bar{x}; \sigma^2 | f_D) = -N \ln 2\pi - \sum_{i=1}^{N} \ln(\rho\gamma_i + \sigma^2) - \sum_{i=1}^{N} \frac{\bar{x}^* \bar{u}_i . \bar{u}_i^* \bar{x}}{\rho\gamma_i + \sigma^2}$$

Nest we invoke the generalization of the Neyman-Fisher factorization theorem, to factor the likelihood function (8) to

$$\Lambda(\bar{x}; \theta) = g(T_1(\bar{x}), T_2(\bar{x}), ..., T_N(\bar{x}), \theta) h(\bar{x}) \tag{8}$$

It is then obvious that the functionals $T_i(x) = |u_i^*.x|^2$ are sufficient for estimating the noise variance σ^2. Due to the orthogonal characteristics of the eigenvectors, it is readily verifiable that the sufficient statistics satisfy the orthogonally principle

$$\mathbf{E}_{\bar{x}}[T_k(x)T_j(x)] = \mathbf{E}_{\bar{x}}[u_k^* x x^* u_j] = u_k^* \sum_{n=1}^{N} \gamma_i u_i . u_i^* u_j = \delta_{kj}(\rho\gamma_k + \sigma^2) \tag{9}$$

Thus, the set $S = \{T_i(x), i = 1, 2, \cdots, N\}$ forms a minimal set of sufficient statistics for estimation of σ^2.

III. Estimator Design

In this section, we consider the estimator design for joint signal and interference power estimation. In our attempt to identify an optimum estimator (in Cramer-Rao sense) for the unknown signal and noise variance, we first show that there does not exist an efficient estimator that attains the CRLB. In doing so, we first compute the gradient of likelihood function with respect to the unknown vector $\kappa = [\sigma^2 \rho]$ as

$$\nabla_\kappa \Lambda(\bar{x}; \kappa) = \begin{bmatrix} -\sum_{i=1}^{N} \frac{1}{(\rho\gamma_i + \sigma^2)} + \sum_{i=1}^{N} \frac{x^* u_i . u_i^* x}{(\rho\gamma_i + \sigma^2)^2} \\ -\sum_{i=1}^{N} \frac{\gamma_i}{(\rho\gamma_i + \sigma^2)} + \sum_{i=1}^{N} \frac{x^* u_i . u_i^* x}{(\rho\gamma_i + \sigma^2)^2} \end{bmatrix} \tag{10}$$

where in the Fisher information matrix is computed as

$$I_\kappa = \begin{bmatrix} \sum_{i=1}^{N} \frac{1}{(\rho\gamma_i + \sigma^2)^2} & \sum_{i=1}^{N} \frac{\gamma_i}{(\rho\gamma_i + \sigma^2)^2} \\ \sum_{i=1}^{N} \frac{\gamma_i}{(\rho\gamma_i + \sigma^2)^2} & \sum_{i=1}^{N} \frac{\gamma_i^2}{(\rho\gamma_i + \sigma^2)^2} \end{bmatrix}$$

Evidently, gradient of the likelihood function can not be factored into the form $\nabla_\kappa \Lambda(\bar{x}; \kappa) = I(\kappa)(\mathbf{g}(\bar{x}) - \kappa)$. Thus, no unbiased estimator would achieve the CRLB bound. The vector parameter CRLB will allow us to place a bound on the variance of each element as $var[\hat{\sigma^2}] \geq I_\kappa^{-1}[1,1]$ and $var[\hat{\rho}] \geq I_\kappa^{-1}[2,2]$.

Now, using the set of sufficient statistics $\{T_j(\bar{x}),\ j = 1, \ldots, N\}$ derived in the previous section, we introduce a general class of estimators based on linear combination of these statistics given by the following equations:

$$\hat{\rho} = \sum_{i=1}^{N} \eta_i T_i(\bar{x}) \tag{11}$$

$$\hat{\sigma^2} = \sum_{i=1}^{N} \beta_i T_i(\bar{x}) \tag{12}$$

Using equation (9), it is easy to see that the above estimators will be unbiased for all possible (ρ, σ^2) combinations, if and only if the coefficients η_i and ξ_i satisfy

$$\sum_{i=1}^{N} \eta_i \gamma_i = 1, \quad \sum_{i=1}^{N} \eta_i = 0, \tag{13}$$

$$\sum_{i=1}^{N} \beta_i \gamma_i = 0, \quad \sum_{i=1}^{N} \beta_i = 1 \tag{14}$$

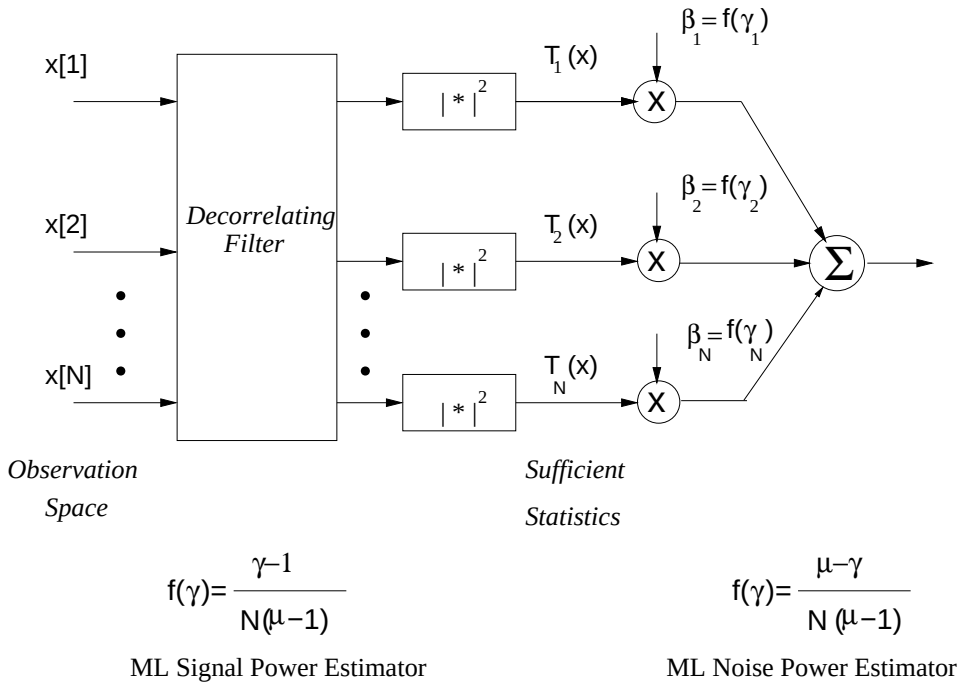


Fig. 2. Estimator Block Diagram

In the complete version of this paper, we show that in the general case the Minimum-Variance Unbiased Estimator (MVUE), if it exists, can not be of the above form. However, we show that two important unbiased estimators, specifically the Maximum Likelihood (ML) and Subspace estimators, do belong to the above-mentioned class.

A. Maximum Likelihood Estimator

Maximum Likelihood Estimator (MLE) has the asymptotic properties of being unbiased, achieving the CRLB and having a Gaussian PDF. Also for high SNRs, MLE achieves the CRLB. MLE obtains the estimate of unknown parameters by maximizing the $\Lambda(\bar{x}, \sigma^2, \rho)$. This joint maximization problem can be formulated as $\nabla_\kappa \Lambda(\bar{x}; \kappa) = 0$. The ML estimates of these two parameters have a closed-form expression as

$$\rho_{ML} = \frac{\sum_{i=1}^{N} T_i(\bar{x})(\gamma_i - 1)}{\mu(N-1)} \tag{15}$$

$$\sigma^2_{ML} = \frac{\sum_{i=1}^{N} T_i(\bar{x})(\mu - \gamma_i)}{\mu(N-1)} \tag{16}$$

where $\mu \triangleq \frac{1}{N}\sum_{i=1}^{N} \gamma_i^2$. Figure (2) shows the structure of the ML estimator.

B. Sub-space Estimator

ML estimation of unknown parameters requires knowledge of both eigenvectors and eigenvalues of the auto-covariance matric. In this section we intend to propose a reduced rank estimator that relaxes this prior information in estimating signal and noise power. Lets define the signal subspace to be the N_s-dimension range space of the autocorrelation matrix R_α. Dimension of signal subspace depends on the doppler parameter and also on the time diversity between successive pilot samples. At low doppler frequencies, the dominant component of the signal lies in rank-1 subspace. As doppler increases, dimension of signal subspace increases accordingly. Similarly, the noise subspace is defined as $(N - N_s)$-dimension null space of the autocorrelation matrix. In another words, noise subspace is the subspace spanned by eigenvectors associated with trivial eigenvalues of R_α ($\{\gamma_i \simeq 0,\ i = [N_{s+1}, \cdots, N]\}$). Thus the observation vector, when projected on noise subspace, has no signal dependent component. As a result, the noise subspace represents projection of noise component onto the null space of R_α. Figure (3) gives an illustration of the signal and noise subspaces. Using this concept, we can estimate the noise power as

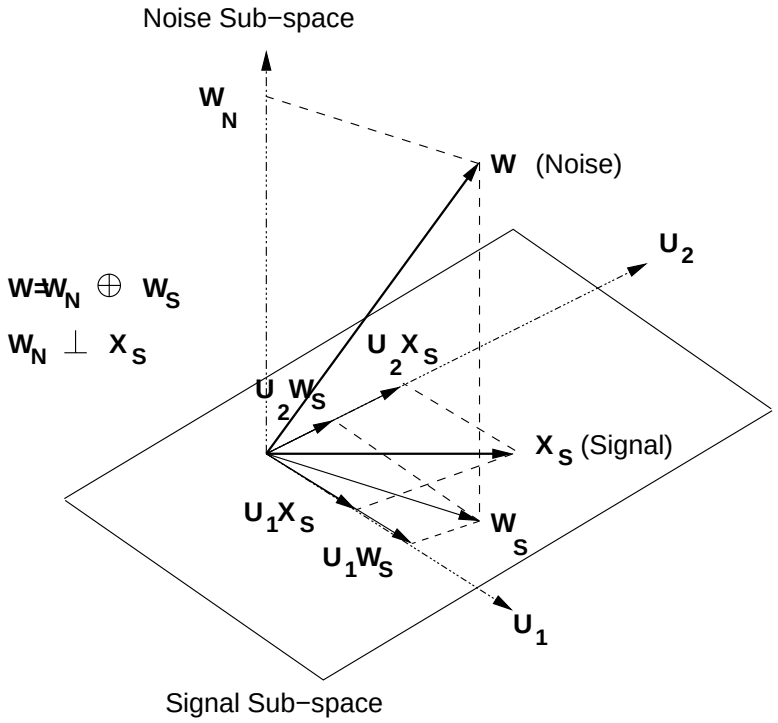


Fig. 3. Signal and Noise Sub-space

$$\sigma^2_{sub} = \frac{1}{N - N_s} \sum_{i=N_s+1}^{N} T_i(\bar{x}) \tag{17}$$

On the other hand, noise has component along the eigenvectors associated with non-trivial eigenvalues of the autocorrelation matrix. Using the orthogonality principle and equation (17), a sub-space signal estimator for signal power is obtained as

$$\rho_{sub} = \frac{1}{N} \sum_{i=1}^{N_s} T_i(\bar{x}) - \frac{N_s}{N(N - N_s)} \sum_{i=N_s+1}^{N} T_i(\bar{x}) \tag{18}$$

Note that both the ML and subspace estimates of signal and noise power require knowledge of the eigenvectors of the underlying channel auto-covariance matrix (R_α). The ML estimator in addition requires the knowledge of the eigenvalues of the same matrix. This requirements would add a constraint for practical implementation of these estimators. However, in the complete version of this paper and for the specific case of 3G Wideband CDMA, we propose an efficient tracking mechanism for estimating signal sub-space in Rayleigh fading (isotropic scattering) scenarios (Jake's model).

The performance of the above estimators can be derived analytically or via simulations. Figure (4) compares the variance of ML and sub-space estimators of noise power for the transmission slot format #0 used in 3G WCDMA (c.f. Figure (1)) and Doppler frequency of $f_D = 200$ Hz. Evidently, as signal power increases, the gap between the CRLB and MLE and Sub-space estimator increases. However, MLE shows less sensitivity in this regard. As noise power increases, ML and Subspace estimators asymptotically reach the CRLB. For ML, this is an inherent property at high SNR. Similarly, when relative energy of noise with respect to signal is increased, projection of signal onto noise sub-space tends to zero.

Figure (5) compares similar experiments for signal power estimates. Note in this case, ML has has relatively uniform distance with respect to CRLB regardless of signal and noise power. When relative power of noise with respect to signal is reduced, sub-space estimator becomes closer to MLE. This is due to the fact that projection of noise on the signal sub-space becomes less effective.

IV. Concluding Remarks

In this paper we have developed an effective class of estimators for joint signal and noise power estimation in Rayleigh fading channels. The estimators rely on the pilot signal tones, although the results can be generalized to blind methods as well. These estimators achieve a performance that approaches asymptotically CRLB. Moreover, the structure of proposed estimators share a commonality that would allow us to investigate their statistical characteristics in a unified framework. A closed form

expression for statistical characteristic of the these estimators, mainly bias and variance, are derived. The proposed class of estimators provide a natural trade-off between complexity and accuracy of the estimators.

A variety of issues remain to be explored in future work. For example, the development of estimator design and analysis techniques specifically optimized for Ricean fading environments can be a valuable resource for general fading environments. More generally, some of the richest directions for future research involve developing techniques for joint estimation of doppler and signal/noise power. Thus, it might be possible to combine the doppler estimator and power estimator into a single efficient estimator.

REFERENCES

[1] N. Freris and T.G. Jeans and P. Taaghol, *Adaptive SIR Estimation in DS-CDMA Caellular Systems Using Kalman Filtering* , IEE Electronics Letter, vol. 37 , pp. 315-317 March. 2001.

[2] S. Gunaratne and P. Taaghol and R. Tafazolli, *Signal Quality Estimation Algorithm* , IEE Electronics Letter, vol. 36 , pp. 1882-1884 Oct. 2000.

[3] T. Jiang and N. D. G. Sidiropoulos and G. B. Giannakis, *Kalman Filtering for Power Estimation in Mobile Communications*, IEEE Trans. Wireless Comms. , vol. 2, No. 1, pp. 151-161, Jan. 2003.

[4] S. Kay, *Fundamentals of Statistical Signal Processing, Vol. I - Estimation Theory* . Prentice Hall, 1993.

[5] D. Pauluzzi and N. Beaulieu, *A Comparison of SNR Estimation Techniques for the AWGN Channel.*, IEEE J. Select. Areas Comms. , vol. 19, No.9, pp. 1697-1705, Sept. 2001.

[6] A. Ramesh and A. Chokalingam and L. B. Milstein, *SNR Estimation in Nakagimi-m Fading with Diversity Combining and Its Application to Turbo Decoding*, IEEE Trans. Comms. , vol. 50, No.11, pp. 1719-1724, Nov. 2002.

[7] M. L. Sim and H. T. Chuah, *Performance evaluation of received power estimation for power control in CDMA systems* , IEEE Veh. Tech. Conference, vol. 3, pp. 1487 -1491, 2002.

[8] G. Stuber, *Principles of Mobile Communication* . 2nd Ed., Kluwer Academic Publishers, 2002.

[9] T.A.. Summers and S. G. Wilson, *SNR Mismatch and Online Estimation in Turbo Decoding*, IEEE Trans. Comms. , vol. 46, No.11, pp. 421-423, Apr. 1998.

[10] J.M.A. Tanskanen and A. Huang and I.O. Hartimo, *Predictive Power Estimators In CDMA Closed Loop Power Control* , Veh. Tech. Conference, vol. 2, pp. 1091-1095, May 1998.

[11] M. Usuda and Y. Ishikawa and S. Onoe, *Optimizing the number of dedicated pilot symbols for forward link in W-CDMA systems* , IEEE Veh. Tech. Conference, vol. 2, pp. 2118-2122, 2000.

[12] M. C. Valenti and B. D. Woerner, *Iterative Channel Estimation and Decoding of Pilot Symbol Assisted Turbo Codes Over Flat-fading Channels*, IEEE Trans. Comm. , vol. 48, No.10, pp. 1681-1691, October 2000.

[13] D. Wong and D. Cox, *Estimating Local Mean Signal Power Level in a Rayleigh Fading Environment.*, IEEE Trans. Veh. Tech. , vol. 48, No.3, pp. 956-959, May 1999.

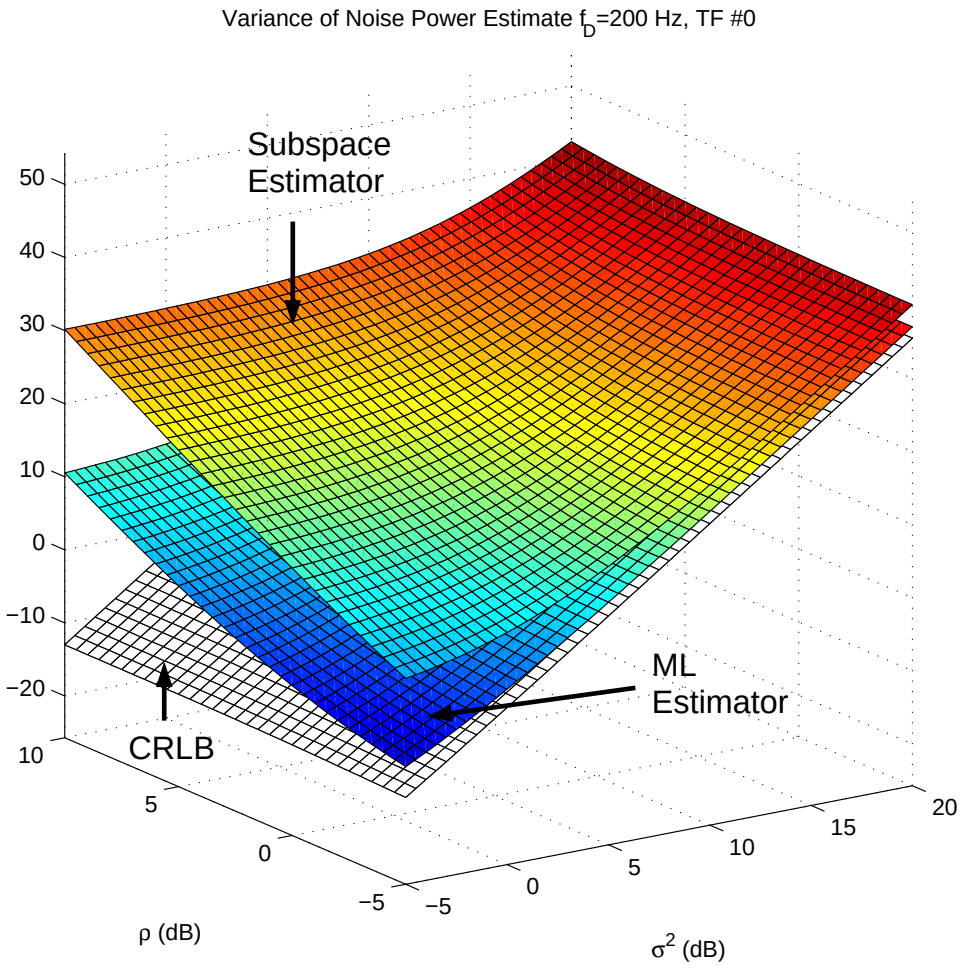


Fig. 4. Comparison of Subspace and ML Noise Power Estimator for TF #0

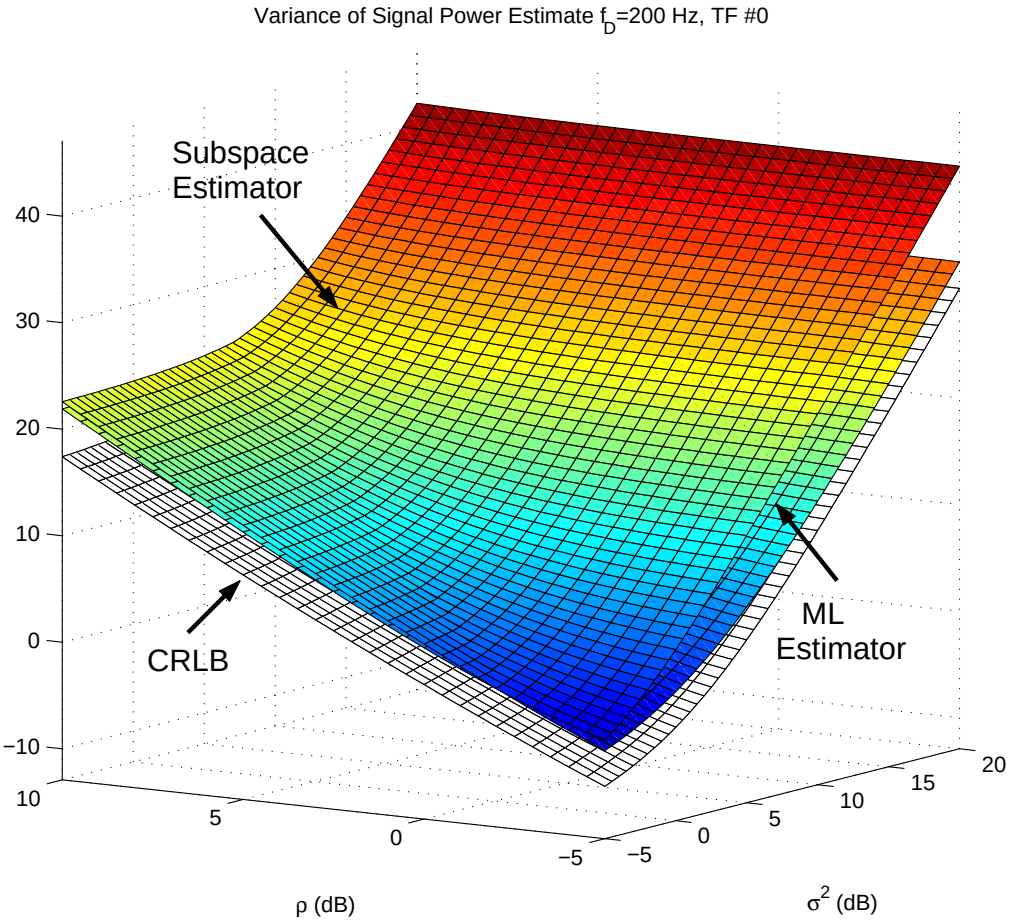


Fig. 5. Comparison of Subspace and ML Signal Power Estimator for TF #0

ONE-STEP REFINEMENT METHOD FOR JOINT CHANNEL ESTIMATION AND TIMING ACQUISITION IN OFDM TRANSMISSION

Marius Sirbu and Visa Koivunen

Signal Processing Laboratory
Helsinki University of Technology
P.O. Box 3000, 02015 HUT
email:{marius,visa}@wooster.hut.£

ABSTRACT

An iterative algorithm for joint time synchronization and channel estimation in a wireless OFDM transmission is presented. The algorithm is based on the knowledge of 2 consecutive identical training symbols. It is well suited for burst transmissions as in wireless LAN (WLAN) applications. The method estimates non-integer time delays and (randomly generated) channel impulse responses. The time delay estimator is robust in the face of carrier frequency offsets while the channel estimator is sensitive only to large carrier frequency offsets. The simulation results indicates good performance at low computational complexity.

1. INTRODUCTION

The multi-carrier systems, and those based on OFDM in particular, have a great potential in offering high data rates. They are very robust to channel frequency selectivity since a frequency selective channel is transformed in a set of parallel ¤at fading channels. Therefore the equalization can be very ef£ciently performed in frequency domain with the inverse of the channel frequency response. Therefore the channel impulse response has to be estimated as well. However, the OFDM based wireless communications (WLAN, DVB and maybe the future 4G mobile wireless networks) are very sensitive to the synchronization errors. Due to the nature of the OFDM signal both time and frequency synchronization have to be performed accurately.

In [1] a blind time delay estimator is proposed in both frequency ¤at and frequency selective channels. It is based on the second order statistics of the received OFDM signal and exploits its cyclostationarity. This algorithm is suitable for non-integer time delays also. The main drawback is the slow time convergence of the statistics which basically limits the algorithm application to continuous transmissions (DVB for example).

This work has been supported by Nokia Foundation, GETA Graduate School and Academy of Finland

A maximum likelihood (ML) time delay estimator is proposed in [2] that exploits the cyclic pre£x in the OFDM signal. The algorithm is designed for integer time delays and the channels are ¤at fading. Algorithms based on the correlation of the received signal are proposed in [3] and [4]. Specially designed training symbols are used in [3], while different smoothing algorithms are used in [4] to reduce the transmission channel effect.

A three step time synchronization procedure is proposed in [5]. In the £rst step the start of the transmission is detected by measuring the received power, then a raw estimated of the time delay is obtained through a correlation technique, and £nally a £ne tuning is performed on the third step. The last two steps are based on known pilot symbols inserted in the transmitted signal.

In this paper, we propose an algorithm to estimate both the time delay and the channel impulse response. The impact of the carrier frequency offset on the estimators behavior is addressed also. We combine the channel estimator presented in [6] with a low complexity delay estimation method in an iterative fashion. The resulting algorithm is able to deliver accurate estimates of the time delay and channel impulse response (after only 2 iterations). For this purpose, only 2 identical consecutive symbols have to be known at the receiver. Therefore the complexity is very low and it may be used with current WLAN standards, for example.

The rest of the paper is organized as follows: the system model is presented in section 2 and the estimation algorithm is derived in section 3. In section 4 simulation results are presented and the conclusions are drawn is section 5.

2. SYSTEM MODEL

Let us assume we have a single user OFDM transmission, where a serial PSK data stream is converted to M parallel

data streams:

$$\mathbf{u}[n] = [u((n-1)M+1), u((n-1)M+2), \ldots, u(nM)]^T \quad (1)$$

The parallel data stream $\mathbf{u}[n]$ is modulated on M orthogonal subcarriers using a Inverse Discrete Fourier Transform (IDFT). In order to avoid the inter-block-interference caused by the channels, a cyclic pre£x is added in front of each OFDM symbol by copying the last P samples of the symbol. The cyclic pre£x length must be longer than the channel impulse response. The signal model may be written as follows:

$$\begin{aligned} \mathbf{s}[n] &= \mathbf{TFu}[n] \quad (2)\\ \mathbf{T} &= \begin{bmatrix} \mathbf{0}_{P\times(M-P)} & \mathbf{I}_P \\ \mathbf{I}_{(M-P)} & \mathbf{0}_{(M-P)\times P} \\ \mathbf{0}_{P\times(M-P)} & \mathbf{I}_P \end{bmatrix} \\ \{\mathbf{F}\}_{mq} &= \frac{1}{\sqrt{M}} e^{j\frac{2\pi mq}{M}}, \quad m,q \in \{0,\ldots,M-1\}, \end{aligned}$$

where $\mathbf{T}$ is the $(M+P)\times M$ matrix which performs the cyclic pre£x addition and $\mathbf{F}$ is the $M\times M$ IDFT matrix. The resulting signals are converted from parallel to serial and then transmitted through the channel with the sampled impulse response $\mathbf{h} = [h(0), \ldots, h(L_h-1)]^T$. The channel impulse response is considered to be time-invariant over a £nite time interval, so called quasi-stationary channel. This condition is in general true for WLAN where the channel impulse response can be considered stationary over a large number of OFDM symbols. Even for applications like mobile communications this property holds but for shorter time periods.

2.1. Asynchronous model

If the receiver and the transmitter are perfectly synchronized (the receiver knows the time instance when the transmission started), the received signal is converted to parallel then the cyclic pre£x is dropped and a DFT operation will restore the original data stream. However in realistic scenarios there are some impairments to deal with. First, the time delay between the transmission and reception has to be estimated in order to properly discard the cyclic pre£x and apply the DFT temporal window correctly. If the delay is smaller than the length of the CP, there is no loss in performance due to the DFT properties, but if the delay is larger than the length of the CP, then the loss in performance is substantial. In order to illustrate this, in Fig. 1 we plotted the loss in performance of the receiver as a function of the timing offset between the transmitter and the receiver in the noiseless case. If the error in the time synchronization estimation is larger than 10% of the sampling period, than the loss in performance is signi£cant. In Fig. 1 we considered perfect channel estimation. The channel and the timing offset have

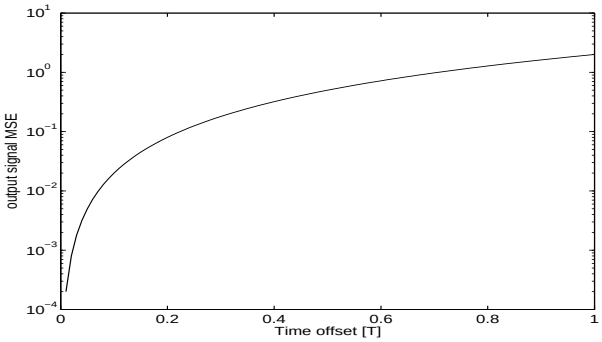


Fig. 1. The performance loss as a function of the timing offset (fraction of the sampling period) between the transmitter and receiver

to be accurately estimated in order to achieve good performance.

We assume that the block level synchronization was performed by measuring the received signal level. Therefore the time offset can be assumed to be smaller than M. Let $\delta \in \mathbb{R}$ be the timing offset between the receiver and transmitter and let p and d be its integer and fractional parts respectively . We can write the received signal as:

$$y(n) = h(n) * s(n-\delta) + w(n) = h(n) * r(n) + w(n), \quad (3)$$

where $r(n)$ is the delayed transmitted signal and $w(n)$ is the additive Gaussian noise and $*$ is the convolution operator. The delayed transmitted signal can be also written as:

$$r(n) = (1-d)s(n-p) + ds(n-p-1) \quad (4)$$

If we stack N ($N = M+P$) consecutive samples in a vector $\mathbf{r}[n] = [r((n-1)N+1), \ldots, r(nN)]^T$, and using the expression (4), we can write:

$$\begin{aligned} \mathbf{r}[n] &= d\left[US(\mathbf{s}[n-1], N-p-1) + DS(\mathbf{s}[n], p+1)\right] + \\ &+(1-d)\left[US(\mathbf{s}[n-1], N-p) + DS(\mathbf{s}[n], p)\right]. \quad (5) \end{aligned}$$

The operators $US(\mathbf{s}, p)$ and $DS(\mathbf{s}, p)$ shift up and down respectively the vector $\mathbf{s}$ with p positions. We can therefore de£ne the up and down shifting matrices $\mathbf{U}(k)$ and $\mathbf{D}(k)$ of sizes $N \times N$:

$$\begin{aligned} \mathbf{U}(k) &= \begin{bmatrix} \mathbf{0}_{(N-k)\times k} & \mathbf{I}_{N-k} \\ & \mathbf{0}_{(k)\times N} \end{bmatrix} \quad (6)\\ \mathbf{D}(k) &= \begin{bmatrix} & \mathbf{0}_{(k)\times k} \\ \mathbf{I}_{N-k} & \mathbf{0}_{(N-k)\times N} \end{bmatrix} \end{aligned}$$

With these de£nitions we can further write:

$$\begin{aligned} &\mathbf{r}[n] = d\left[\mathbf{U}(N-p-1)\mathbf{s}[n-1] + \mathbf{D}(p+1)\mathbf{s}[n]\right] + \\ &+(1-d)\left[\mathbf{U}(N-p)\mathbf{s}[n-1] + \mathbf{D}(p)\mathbf{s}[n]\right]; \\ &\mathbf{r}[n] = \left[d\mathbf{U}(N-p-1) + (1-d)\mathbf{U}(N-p)\right]\mathbf{s}[n-1] + \\ &+\left[d\mathbf{D}(p+1) + (1-d)\mathbf{D}(p)\right]\mathbf{s}[n]; \\ &\mathbf{r}[n] = \mathbf{B}\tilde{\mathbf{s}}[n] = \mathbf{B}(\mathbf{I}_2 \otimes \mathbf{TF})\tilde{\mathbf{u}}[n], \quad (7) \end{aligned}$$

where $\otimes$ is the Kroneker product, $\mathbf{I}_2$ is the 2×2 identity matrix and:

$$\mathbf{B} = \Big[d\mathbf{U}(N-p-1) + (1-d)\mathbf{U}(N-p) \ \vdots$$
$$\vdots \ d\mathbf{D}(p+1) + (1-d)\mathbf{D}(p) \Big];$$
$$\tilde{\mathbf{s}}[n] = \left[\mathbf{s}^T[n-1] \ \ \mathbf{s}^T[n]\right]^T;$$
$$\tilde{\mathbf{u}}[n] = \left[\mathbf{u}^T[n-1] \ \ \mathbf{u}^T[n]\right]^T. \quad (8)$$

Our goal in this paper is to jointly estimate the channel impulse response $\mathbf{h}$ and the delay δ.

3. ALGORITHM

The channel estimation scheme proposed in [6] is based on the IEEE 802.11 standard for perfectly synchronized transmission. We extend this scheme for the estimation of the time delays also. In the IEEE 802.11 standard, two known and equal training symbols are available, therefore we can use them to estimate the channel. An optimal solution is to estimate the channel for all possible time delays δ, and choose the best estimate based on a MMSE criterion . But such a solution is far too complex. We propose a method to iteratively estimate the time delay and channel impulse response.

Let us assume the received serial signal is passed through a serial to parallel converter and the cyclic pre£x is removed. The received signal vector after CP removal is:

$$\mathbf{y}_p[n] = [y((n-1)M + nP + 1), \ldots, y(n(M+P))]^T \quad (9)$$

It can be also written as:

$$\mathbf{y}_p[n] = \mathbf{R}[n]\mathbf{h} + \mathbf{w}[n], \quad (10)$$

where $\mathbf{w}[n]$ is the noise vector and $\mathbf{R}[n]$ is the signal convolution matrix of size $M \times L_h$, with a special structure due to the cyclic pre£x:

$$\mathbf{R}[n] = \begin{bmatrix} r((n-1)M+nP+1) & r(n(M+P)) \\ r((n-1)M+nP+2) & r((n-1)M+nP+1) \\ \vdots & \vdots \\ r(n(M+P)) & r(n(M+P)-1) \end{bmatrix}$$
$$\begin{matrix} \ldots & r(n(M+P)-L_h+2) \\ \ldots & r(n(M+P)-L_h+3) \\ & \vdots \\ \ldots & r(n(M+P)-L_h+1) \end{matrix} \Bigg] \quad (11)$$

The received vectors $\mathbf{y}[1]$ and $\mathbf{y}[2]$ are the result of the transmission of the same training OFDM symbol $\mathbf{r}[1]$.

The £rst step of our algorithm is to consider an arbitrary valued vector as the initial estimate of the channel impulse response $\hat{\mathbf{h}}_0$. We generate the convolution matrices $\mathbf{R}_\tau[n]$ corresponding to the known training symbols and the possible integer delays $\tau \in \{0, \ldots, M-1\}$. A £rst initial estimate of the time delay can be found by solving:

$$\mathcal{J}(\tau) = \left(\mathbf{y}[2] - \mathbf{R}_\tau[2]\hat{\mathbf{h}}_o^T\right)^H \left(\mathbf{y}[2] - \mathbf{R}_\tau[2]\hat{\mathbf{h}}_o^T\right)$$
$$\hat{\delta}_1 = \arg \min_{\tau \in \{0,\ldots,M-1\}} \mathcal{J}(\tau) \quad (12)$$

Having the initial estimate $\hat{\delta}_1$, we can estimate the channel impulse response as follows:

$$\hat{\mathbf{h}}_1 = \mathbf{R}^{\#}_{\hat{\delta}_1}[2] \frac{\mathbf{y}[1] + \mathbf{y}[2]}{2}, \quad (13)$$

where the operator $(.)^{\#}$ stands for the pseudo-inverse, therefore the size of $\mathbf{R}^{\#}_{\hat{\delta}_1}[2]$ is $L_h \times M$.

We have found so far a raw estimate of the timing offset $\hat{\delta}_1$ and an estimate of the channel impulse response, $\hat{\mathbf{h}}_1$.

Using the same received data we can re£ne the estimated delay by searching for the minimum of the same quadratic cost function in (12) but searching on the interval $\{\hat{\delta}_1 - G, \ldots, \hat{\delta}_1 + G\}$ with a smaller grid, for example 0.1.

$$\hat{\delta}_2 = \arg \min_{\tau \in \{\hat{\delta}_1 - G : 0.1 : \hat{\delta}_1 + G\}} \mathcal{J}(\tau). \quad (14)$$

The value G have to be set such that the second searching interval, $\{\hat{\delta}_1 - G, \ldots, \hat{\delta}_1 + G\}$, includes the true delay δ, otherwise the method fails. From experimental simulations we observed that after the £rst iteration, the estimated delay is in the range $\{\delta + 1, \ldots, \delta - 1\}$, therefore we set $G = 2$ to ensure that the algorithm does not fail.

In Fig. 2 we plotted the cost function $\mathcal{J}$ for two iterations.

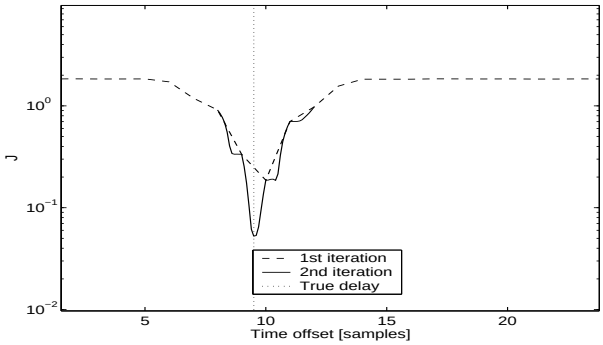


Fig. 2. The cost function $\mathcal{J}$ for 2 iterations.

An improved channel estimator can be found by using the new estimated value $\hat{\delta}_2$ in equation (13).

$$\hat{\mathbf{h}}_2 = \mathbf{R}^{\#}_{\hat{\delta}_2}[2] \frac{\mathbf{y}[1] + \mathbf{y}[2]}{2}, \quad (15)$$

Therefore in 2 iterations on the same received data we are able to estimate both the timing delay and the channel impulse response. The algorithm may be summarized as follows:

1. *Raw estimation*
   - *Consider an arbitrary valued vector as the initial estimate of the channel impulse response,* $\hat{\mathbf{h}}_0$*, and compute the £rst estimate of the time delay* $\hat{\delta}_1$*, using (12).*
   - *By using* $\hat{\delta}_1$ *and (13) re£ne the channel impulse response estimate,* $\hat{\mathbf{h}}_1$.
2. *Re£nement*
   - *Using* $\hat{\mathbf{h}}_1$ *and a smaller grid search in (12), re£ne the time delay estimate,* $\hat{\delta}_2$.
   - *Re£ne the channel impulse response estimate using the time delay estimate* $\hat{\delta}_2$ *in (13).*

3.1. Discussion

So far we have assumed that the carrier frequency offset is zero. However in a realistic scenario carrier frequency offset exists and it has to be compensated for. Otherwise a severe degradation in the receiver performance is expected due to the inter-carrier-interference. Therefore each received OFDM symbol is multiplied by a factor $e^{j2\pi\Delta f n}$ where $\Delta f \in [-0.5, 0.5)$ is the normalized frequency offset between the transmitter carrier frequency and receiver carrier frequency.

The cost function $\mathcal{J}$ computed in (12) is robust to the frequency offset Δf due to its quadratic form. However, the channel estimator performance may be highly degraded because of large frequency offsets. In Fig. 3 we depicted the time delay estimation error and the channel estimate MSE for different frequency offsets. For small values of the normalized frequency offset (smaller than 0.01 of the subcarrier spacing) both estimators performs very well. For large frequency offset the channel estimator performance degrades very quickly while the time delay estimator still performs well.

The effect of carrier frequency offset may be mitigated by including a frequency offset estimator in our algorithm. Even for large frequency offset the time delay is well estimated after the second iteration. Therefore after the second iteration the algorithm proposed in [7] may be used to estimate the frequency offset. This algorithm employs the same system model setup with two identical training symbols. A third iteration is necessary in this case to re£ne the estimate of the channel impulse response.

In the following we summarize the frequency offset estimator. The n-th received OFDM symbol with the frequency offset can be written as $\mathbf{y}_{offset}[n] = \mathbf{y}_P[n]e^{-j2\pi\Delta f nT}$,

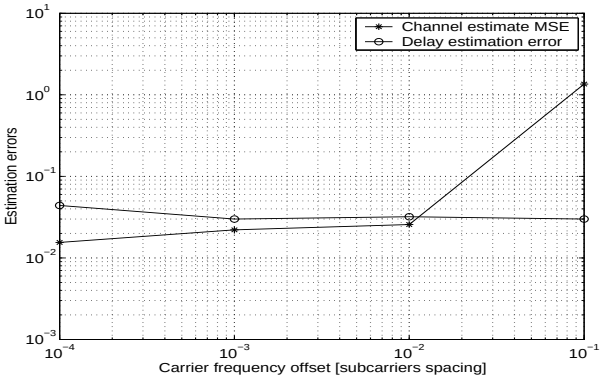


Fig. 3. The estimator performance for imperfect carrier offset compensation ($E_b/N_0 = 10$dB).

where Δf is the normalized frequency offset. Therefore if two OFDM symbols are known and equal, we can write for the corresponding received symbols, in the noiseless case:

$$z = \mathbf{y}^H_{offset}[n]\mathbf{y}_{offset}[n+1] = \mathbf{y}^H_P[n]\mathbf{y}_P[n]e^{-j2\pi\Delta fT} = \\ = \| \mathbf{y}[n] \|^2_F \, e^{-j2\pi\Delta fT} \quad (16)$$

where $\| \, . \, \|^2_F$ is the squared Frobenius norm. The frequency offset can be determined with:

$$\hat{\Delta} f = -\frac{1}{2\pi T} \arg z \quad (17)$$

The frequency offset estimator performance is degraded in the presence of additive noise.

4. SIMULATION RESULTS

In this section we present the simulation results to illustrate the performance of the algorithm. We use a similar packet data transmission as in IEEE 801.11 standard. Two known OFDM symbols are transmitted in the £rst data packet (burst) for channel estimation and timing acquisition purposes. Then the other packets contain only information symbols. A QPSK signal is modulated on $M = 64$ subcarriers and a CP of length 6 is added. The channel impulse response used in simulations has 4 taps, is randomly generated and its norm is unity. The channel is considered to be time-invariant over the observation interval. The time delay is generated randomly in the interval $[0, M)$. The simulation results are averaged over 100 independent runs.

We will demonstrate that the algorithm converges after 2 iterations on the received training symbols. In Fig. 4 the estimation error of the time delay and the mean square error (MSE) of the channel estimate are presented in a scenario where 10 iterations have been performed. The results are that expected, after 2 iterations the estimates quality does

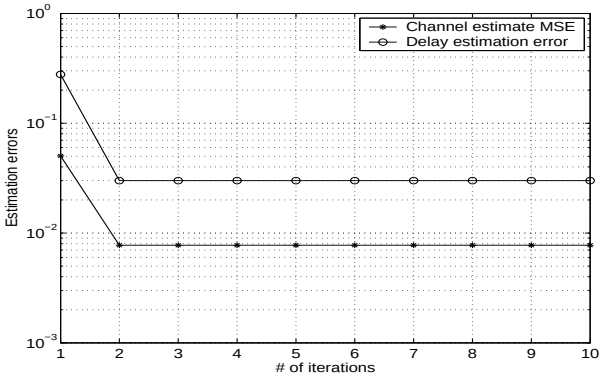


Fig. 4. The estimation errors for 10 iterations ($E_b/N_0 = 10$dB).

not improve anymore. Therefore there is no need to increase the number of iterations over 2.

We also observe that the time delay estimation error is well bellow 10% of the sampling period. This will ensure a very small deterioration in the receiver performance (see Fig.1).

A good time delay estimator has to be as insensitive as possible to the additive noise level. In Fig.5 the bias and variance of the time delay estimator after 2 iterations are presented for different noise levels. The estimator performance is almost constant regardless of the signal to noise ratio. The bias and the variance of the delay estimator where computed as:

$$\text{bias} = \left| \delta - \frac{1}{N_r} \sum_{k=1}^{N_r} \hat{\delta}_2(k) \right| \tag{18}$$

$$\text{var} = \frac{1}{N_r} \sum_{k=1}^{N_r} [\hat{\delta}_2(k) - \delta]^2 - bias^2, \tag{19}$$

where N_r is the number of independent runs.

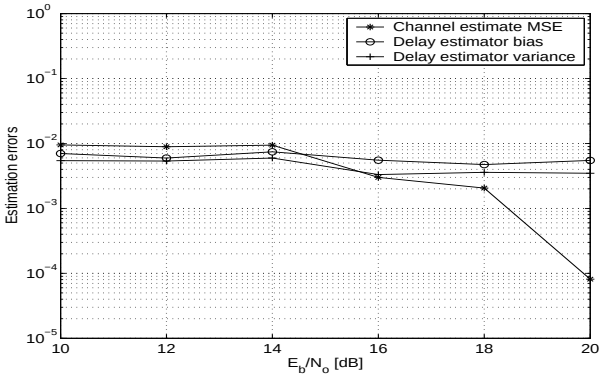


Fig. 5. The bias and variance of the time delay estimator and the MSE of the channel estimate for different E_b/N_0.

In Fig. 5, the MSE of the channel estimator is also plotted. Unlike the time delay estimator the channel estimator performance is signi£cantly improved as the E_b/N_0 increases. This is an expected behavior since the channel estimate is directly in¤uenced by the additive noise level (see equation (13)).

5. CONCLUSIONS

We proposed a method for time synchronization and channel estimation in OFDM transmissions. Two known and equal OFDM symbols are necessary in the proposed scheme. The algorithm works in a iterative fashion on the same received data. The time delay estimator is very robust to the carrier frequency offsets while the channel estimator performance is degraded by large frequency offsets. Both estimators have a reliable behavior in the presence of additive noise.

6. REFERENCES

[1] H. Bolcskei, "Blind estimation of symbol timing and carrier frequency offset in wireless OFDM systems," *IEEE Trans. Commun.*, vol. 49, no. 6, pp. 988–999, June 2001.

[2] D. Landstrom, S. Wilson, J.-J. van de Beek, P. Odling, and P. Borjesson, "Symbol time offset estimation in coherent OFDM systems," *IEEE Trans. Commun.*, vol. 50, no. 4, pp. 545–549, April 2002.

[3] T. Schmidl and D. Cox, "Robust frequency and timing synchronization for OFDM," *IEEE Trans. Commun.*, vol. 45, no. 12, pp. 1613–1621, December 1997.

[4] M.-H. Hsieh and C.-H. Wei, "A low-complexity frame synchronization and frequency offset compensation scheme for OFDM systems over fading channels," *IEEE Trans. Veh. Technol.*, vol. 48, no. 5, pp. 1596–1609, September 1999.

[5] W. Warner and C. Leung, "OFDM/FM frame synchronization for mobile radio data communication," *IEEE Trans. Veh. Technol.*, vol. 42, no. 3, pp. 302–313, August 1993.

[6] J. Heiskala and J. Terry, *OFDM Wireless LANs: A theoretical and practical guide*. Sams Publishing, 2002, ch. 2.

[7] J.-J. van de Beek, M. Sandell, and P. Borjesson, "Ml estimation of time and frequency offset in OFDM systems," *IEEE Trans. Signal Proc.*, vol. 45, no. 7, pp. 1800–1805, July 1997.

Computational Intelligence Techniques for Overcoming Co-Channel Interference in Mobile Cellular Networks

Arnaud Olivier & Amir Hussain
Department of Computing Science and Mathematics
University of Stirling, Stirling FK9 4LA, SCOTLAND
corresponding Author's E-mail: aov@cs.stir.ac.uk, ahu@cs.stir.ac.uk

Abstract

In this paper, non-linear adaptive Feedforward and novel Decision Feedback equalizers based on Support Vector Machine (SVM) and Wavenets are applied to the problem of adaptive equalization in the presence of Inter-Symbol Interference, Additive White Gaussian Noise, and Co-Channel Interference. A realistic severe amplitude distorted co-channel system is used as a case study to illustrate the superior Bit Error State performance of the proposed computational intelligence based adaptive equalizers compared to linear and non-linear equalizers.

1 Introduction

Adaptive equalization is known to be an important technique for combating distortion and Inter-Symbol Interference (ISI) in communication channels. However, many communication systems are also impaired by what is known as co-channel interference (CCI). Many digital communications systems such as cellular radio and dual polarized micro-wave radio, for example, employ frequency reusage and often exhibit performance limitation due to co-channel interference [1]. Frequency reuse is referred to the employment of radio channels on the same carrier frequency to cover different areas or cells situated sufficiently apart from one other, and allows cellular radio systems to handle far more simultaneous calls than the total number of allocated channel frequencies. Signals for co-channel cells (i.e. cells of the same channel frequency) will however interfere with each other thus requiring the use of adaptive equalizers in these communications systems for reliable data transmission.

Two basic categories of adaptive equalizers exist, namely the sequence estimation and symbol decision equalizers. The optimal sequence estimation equalizer is the Maximum Likelihood sequence estimator (MLSE) which provides the best attainable performance in combating channel ISI and Additive White Gaussian Noise (AWGN) at the expense of very high computational complexity and deferring decisions. The MLSE is however much less effective in dealing with Co-Channel Interference. It can best treat the unknown interfering signal as an additional colored noise since it is very difficult to derive the likelihood function for the non-gaussian interfering signals to enable them to be explicitly distinguishable from the gaussian noise. Most of the equalization applications today employ equalizers that operate symbol-by-symbol. Symbol decision equalizers can be further qualified into two categories namely, the direct-modelling equalizers in which the channel model is identified explicitly, and the indirect-modelling equalizers which recover the transmitted symbols by directly filtering the channel observations, usually using the Linear Transversal Equalizer, without estimating a channel model explicitly. The indirect-modelling approach is by far most widely used and it is considered in the present study in the context of Co-Channel Interference.

It is well-known that the LTE does not achieve the full performance potential of the symbol-decision equalizer structure for combating channel

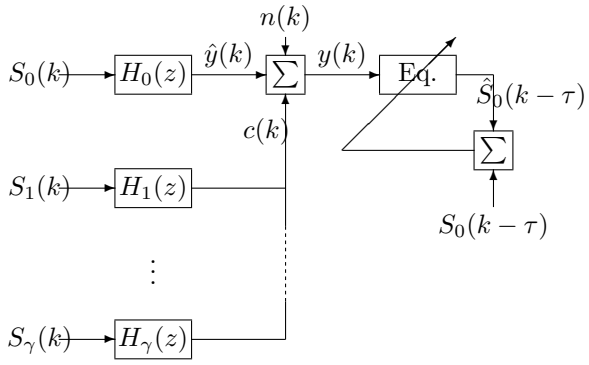


Figure 1: System model

ISI in the presence of additive white gaussian noise, as it fails to make use of the fact that transmitted digital symbols can only take on finite value; and it also suffers from the problem of noise enhancement [1]. Better performance can be achieved form the same information contained in the equalizer inputs, if more complex filtering methods are employed.

The current work is an application of two new computational intelligence tools, namely the Wavelet Neural Network ([2, 4, 5]), and the Support Vector Machine ([6]). The two novel methods are applied to the problem of combating co-channel interference, without any apriori knowledge of the channel / co-channel orders.

This paper is organized as follow:

In section 2, the discrete time model of the digital communication system is presented. Section 3 introduces the two new equalizer structures and their associated learning algorithms. Simulation results are presented in section 4, where the equalizers are applied to a realistic co-channel system and their BER performance characteristics compared. Finally, section 5 presents some concluding remarks.

2 System model

Following [1], the discrete time model of the data transmission system considered in this paper is shown in Figure 1. In this model, $H_0(z)$ is the dispersive channel transfer function and $H_1(z) \dots H_\gamma(z)$ represent the interfering co-channels. All channels are modelled by Finite Impulse Response (FIR) filters as:

$$H_i(z) = \sum_{j=0}^{l_i} h_{ij} z^{-j} \;\; i = 0, 1, \dots, \gamma \tag{1}$$

In Figure 1, $s_0(k)$ represent the transmitted data (which is know during the equalizer training phase) and $s_i(k), i = 1, \dots, \gamma$ are unknown interfering data sequences. All $s_i(k), i = 0, \dots, \gamma$ are assumed to be equiprobable and bipolar independent identically distributed (iid) and the output from the co-channels $c(k)$ are corrupted by Additive White Gaussian Noise (AWGN) $n(k)$ of zero mean and variance σ_n^2. All $s_i(k), i = 0, 1, \dots, \gamma$ are assumed to be uncorrelated with $n(k)$. The overall channel observation can thus be written as:

$$y(k) = \hat{y}(k) + c(k) + n(k) \tag{2}$$

where

$$\hat{y}(k) = \sum_{j=0}^{l_0} h_{0j} s_0(k-j) \tag{3}$$

where l_0 is order of the distorting channel; and

$$c(k) = \sum_{i=1}^{\gamma} \sum_{j=0}^{l_i} h_{ij} s_i(k-j) \tag{4}$$

where $l_i, i = 1, \dots, \gamma$ is the order of the $i-th$ interfering co-channel.

If $E\{\hat{y}^2(k)\} = \sigma_s^2$ (where $E\{.\}$ is the expectation operator) and $E\{c^2(k)\} = \sigma_c^2$; then the following expressions can be defined:

The Signal to Noise Ratio (SNR) given by:

$$SNR = \frac{\sigma_s^2}{\sigma_n^2} \tag{5}$$

The Signal to Interference Ratio (SIR) defined as:

$$SIR = \frac{\sigma_s^2}{\sigma_c^2} \tag{6}$$

And finally the Signal to Interference to Noise Ratio (SINR) given by:

$$SINR = \frac{\sigma_s^2}{(\sigma_n^2 + \sigma_c^2)} \tag{7}$$

The task of any indirect-modelling equalizer is: given the overall channel observation $y(k)$, estimate the transmitted data $s_0(k)$. The symbol decision equalizer at any sample instant k processes the n

most recent channel observations, and makes a decision $\hat{s}_0(k-\tau)$ regarding the symbol transmitted at $k-\tau$, where integer n and τ are referred to as the equalizer order and delay respectively. For decision Feedback structures, a Feedback order m is also added. Thus, $s_0(k-\tau)$ is estimated from the n most recent channel observations, and the m past decisions of the equalizer.

During the training period of most adaptive equalization systems (including cellular mobile radio), the reference desired signal $s_0(k-\tau)$ which is to be re-constructed, is available, whereas the other interfering signals $s_i(k), i = 1, \ldots, \gamma$ are not know.

3 The computational intelligence based structures and their Learning Algorithms

The following section will describe the two proposed structures, and their learning algorithms.

In all cases, a process with N_i inputs, and a single output is considered. During the training phase of the equalizer, a set of N training pairs is used:

$$(\mathbf{x}^n, y^n), n = 1, \ldots, N \tag{8}$$

where:

$$\mathbf{x}^n = [x_1^n, x_2^n, \ldots, x_{N_i}^n] \tag{9}$$

3.1 Support Vector Machine

Support Vector Machines [6, 7, 8, 9, 10, 11] are a powerful approach for solving classification and regression problems. The formulation of SVM embodies the structural risk minimization (SRM) principle. SRM minimize an upper bound on the expected risk, as opposed to the traditional empirical risk minimization that minimizes the error on the training data. This difference equips SVM with a great ability to generalize. For two-group classification problems, generalization ability is optimized by maximizing the margin between the decision hyperplane and the training data, and the solution is obtained as a set of sparse supports vectors, which lie on the margin boundary and summarize the information required to separate the data.

This is realized by solving the following quadratic problem (a complete derivation is given in [6]).

$$W(\alpha) = \sum_{n=1}^{N} \alpha^n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha^n \alpha^m y^n y^m K(\mathbf{x^n}, \mathbf{x^m}) \tag{10}$$

with constraints:

$$0 \leq \alpha^n \leq C, n = 1, \ldots, N \tag{11}$$

$$\sum_{n=1}^{N} \alpha^n y^n = 0 \tag{12}$$

where $K(.,.)$ is the Kernel function, C is a given value representing the penalty for misclassified patterns, $(\mathbf{x}^n, y^n), n = 1, \ldots, N$ are the training pairs defined in Equation (8), and $\alpha^n, n = 1, \ldots, N$ are the Lagrange multipliers.

The Kernel $K(.,.)$ correspond to a inner product of vectors in a higher dimensional feature space if and only if Mercer's condition is met ([9, 11]). Some common kernels meeting this condition are shown in Table 1.

Linear	$K(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1.\mathbf{x}_2$
Polynomial	$K(\mathbf{x}_1, \mathbf{x}_2) = (\mathbf{x}_1.\mathbf{x}_2 + 1)^d$
RBF	$K(\mathbf{x}_1, \mathbf{x}_2) = \exp(-\|\|\mathbf{x}_1 - \mathbf{x}_2\|\|^2)/2\sigma^2$
Sigmoidal	$K(\mathbf{x}_1, \mathbf{x}_2) = \tanh(\kappa \mathbf{x}_1.\mathbf{x}_2 - \delta)$

Table 1: Some Kernel functions for the SVM

In this study, only the polynomial kernel will be considered. This kernel is more efficient for real-time application than the other kernels, which require computation of the exponential function, and it allows to map the input vector into a higher dimensional space, which may result in better classification. This assertion will be experimented later.

Solving Equation (10) with constraints (11, 12) determine the Lagrange multipliers. The training inputs $\mathbf{x}^n, n = 1, \ldots, N$ which have non-zero Lagrange multipliers are called Support Vectors.

Then, during the testing phase, the classification is realized with a hard classifier:

$$y = f(\mathbf{x}) = sign(\sum_{n \in SV}^{N} \alpha^n K(\mathbf{x}^n, \mathbf{x}) + b) \tag{13}$$

where :

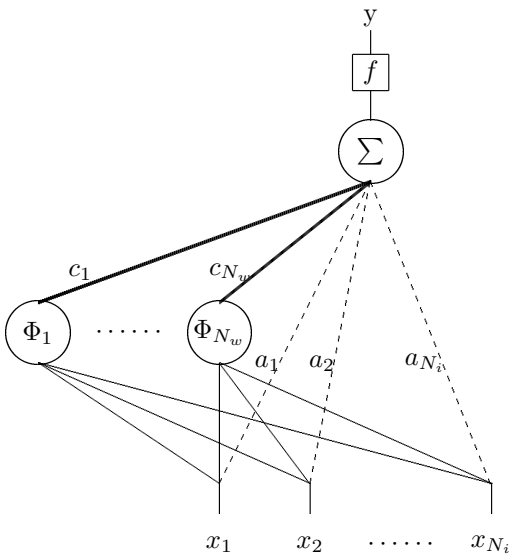


Figure 2: A wavelet network

$$b = \frac{1}{|SV|} \sum_{n \in SV} (y^n - \sum_{m=1}^{N} \alpha^m y^m K(\mathbf{x}^n, \mathbf{x}^m)) \quad (14)$$

SV is the set containing all the indices of the support vectors, and $|SV|$ denote the cardinality of the set SV.

Following Equation (13), all input vectors with zero-valued Lagrange multipliers can be pruned without any loss of information.

Mote that the values of C (misclassification penalty) and d (polynomial order of the kernel function) are determined on a trail and error basis.

3.2 Wavenet

The theory of wavelets was first proposed in the field of multiresolution analysis; among others, it has been applied to image and signal processing ([3]). A family of wavelets is constructed by translations and dilations performed on a single fixed function called the mother wavelet. A wavelet ϕ_j is derived from its mother wavelet ϕ by:

$$\phi_j(z) = \phi(\frac{x - m_j}{d_j}) \quad (15)$$

where its translation factor m_j and its dilation factor d_j are real numbers ($d_j > 0$).

For classification purposes, different approaches exist ([2]), but only one will be considered in this paper, namely, we will consider a weighted sum of wavelets functions whose parameters m_j and d_j are adjustable real numbers, which will be trained together with the weights

Wavelets can thus be considered as a family of parameterized nonlinear functions which can be used for nonlinear classification with their parameters being estimated through "training".

In the case of a problem with N_i inputs, multidimensional wavelets must be considered. The simplest, most frequent choice (as in [2]) is that of separable wavelets, i.e. the product of N_i monodimensional wavelets of each input:

$$\Phi_j(\mathbf{x}) = \prod_{k=1}^{N_i} \phi(z_{jk}) \; with \; z_{jk} = \frac{x_k - m_{jk}}{d_{jk}} \quad (16)$$

where m_j and d_j are the translation and dilatation vectors.

We consider wavelet networks of the form:

$$\psi(\mathbf{x}) = \sum_{j=1}^{N_w} c_j \Phi_j(\mathbf{x}) + a_0 + \sum_{k=1}^{N_i} a_k x_k \quad (17)$$

Equation (17) can be viewed as a network with N_i inputs, a layer of N_w wavelets of dimension N_i, a bias term, and a linear output neuron. When linear terms are expected to play an important role in the model, it is customary to have additional direct connections from inputs to outputs, since there is no point in using wavelets for reconstructing linear terms. Besides, as in neural networks, different transfer functions f can be used, under the constraint of being differentiable everywhere. Examples are shown in Table 2.

Linear	$y = f(\psi(\mathbf{x})) = \psi(\mathbf{x})$
Sigmoidal	$y = f(\psi(\mathbf{x})) = \tanh(\psi(\mathbf{x}))$

Table 2: Some examples of transfer function for the Wavenet

Such a network is named Wavelet Neural Network, or simply a Wavenet, and is shown in Figure 2.

The training of a Wavenet is based on the minimization of the following quadratic cost function, using the training pairs defined in Equation (8):

$$J(\theta) = \frac{1}{2}(y^n - \bar{y}^n)^2 = \frac{1}{2}(e^n)^2 \tag{18}$$

where

$$\theta = \{m_{jk}, d_{jk}, c_j, a_k, a_0\} \tag{19}$$

with $j = 1, \ldots, N_w$ *and* $k = 1, \ldots, N_i$, being all the free parameters of the Wavenet, and $\bar{y}^n = f(\psi(\mathbf{x}^n))$ the calculated Wavenet output.

The minimization is performed by iterative gradient-based methods. The partial derivative of the cost function with respect to θ is:

$$\frac{\partial J}{\partial \theta} = -e^n \frac{\partial \bar{y}^n}{\partial \theta} = -e^n \frac{\partial \bar{y}^n}{\partial \psi^n} \frac{\partial \psi^n}{\partial \theta} \tag{20}$$

Where the term $\frac{\partial y^n}{\partial \psi^n}$ is the derivative of the transfer function f with respect to his parameter:

$$\left.\frac{\partial \bar{y}^n}{\partial \psi^n}\right|_{x=x^n} = f'(\bar{y}^n) \tag{21}$$

And the different components of $\frac{\partial \psi^n}{\partial \theta}$ are:

- parameter a_0:

$$\frac{\partial \psi^n}{\partial a_0} = 1 \tag{22}$$

- direct connection parameters:

$$\frac{\partial \psi^n}{\partial a_k} = x_k^n, k = 1, \ldots, N_i \tag{23}$$

- weights:

$$\frac{\partial \psi^n}{\partial c_j} = \Phi(\mathbf{x}^n), j = 1, \ldots, N_w \tag{24}$$

- translations:

$$\frac{\partial \psi^n}{\partial m_{jk}} = -\frac{c_j}{d_{jk}} \left.\frac{\partial \Phi_j}{\partial z_{jk}}\right|_{x=x^n} \tag{25}$$

with $k = 1, \ldots, N_i$ *and* $j = 1, \ldots, N_w$, and

$$\left.\frac{\partial \Phi_j}{\partial z_{jk}}\right|_{x=x^n} = \phi(z_{j1}^n) \ldots \phi'(z_{jk}^n) \ldots \phi(z_{jN_i}^n) \tag{26}$$

Nw \ f	linear	sigmoidal
2	-13.2563	-13.9058
3	-12.336	-14.5442
4	-12.067	-14.1885
5	-11.9797	-13.696
6	-12.6049	-13.8453
7	-12.6021	-14.3709
8	-11.884	-13.798

Table 3: BER results with different values of Nw and two different transfer functions for the Wavenet-based equalizer (transversal case)

where $\phi'(z_{jk}^n)$ is the value of the derivative of the scalar mother wavelet at point z_{jk}^n:

$$\phi'(z_{jk}^n) = \left.\frac{d\phi(z)}{dz}\right|_{z_{jk}^n} \tag{27}$$

- dilatations:

$$\frac{\partial \psi^n}{\partial d_{jk}} = -\frac{c_j}{d_{jk}} z_{jk}^n \left.\frac{\partial \Phi_j}{\partial z_{jk}^n}\right|_{x=x^n} \tag{28}$$

with $k = 1, \ldots, N_i$ *and* $j = 1, \ldots, N_w$

At each iteration, and at each pattern, the parameters are modified using the gradient (20), according to:

$$\Delta\theta = -\mu \frac{\partial J}{\partial \theta}, 0 < \mu < 1 \tag{29}$$

where μ is real valued and represents the learning rate.

In this study, the first Gaussian derivative was used as the mother wavelet ($f(z) = -z * exp(-z.^2/2)$). The choice of the size of the hidden layer (e.g. N_w), and of the type of the transfer function f, will be made on a trail and error basis.

4 Simulation Results and Comparison with other equalizers

As in [1], a severe amplitude distorted co-channel involving one interfering co-channel, rep-

Nw \ f	linear	sigmoidal
2	-14.3224	-17.5446
3	-14.1691	-17.7201
4	-15.498	-17.3974
5	-15.891	-16.9293
6	-13.8465	-15.4014
7	-15.4157	-17.2444
8	-14.6782	-16.908

Table 4: BER results with different values of Nw and two different transfer functions for the Wavenet-based equalizer (feedback case)

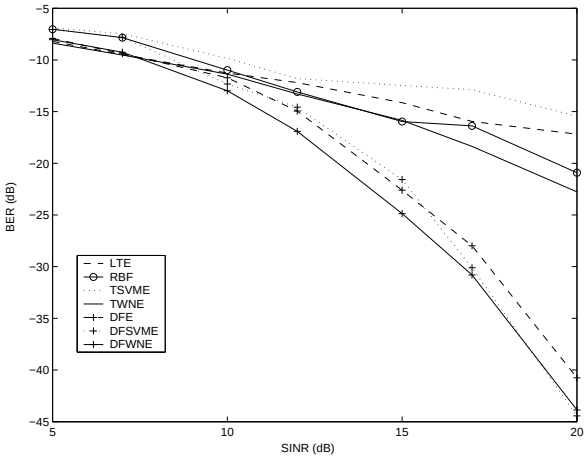


Figure 3: Comparison between the DFSVME, DFWNE, LTE, DFE, and RBF in case 1

C \ d	2	3	4	5
1	-13.2	-12.0	-10.9	-12.2
2	-12.7	-12.1	-11.4	-10.4
5	-11.4	-11.6	-10.4	-10.5
10	-12.4	-11.7	-11.5	-11.2
50	-11.3	-12.6	-10.8	-10.7
∞	-10.9	-11.3	-10.5	-10.8

Table 5: BER results with different values of C and d for the SVM-based equalizer (Transversal case)

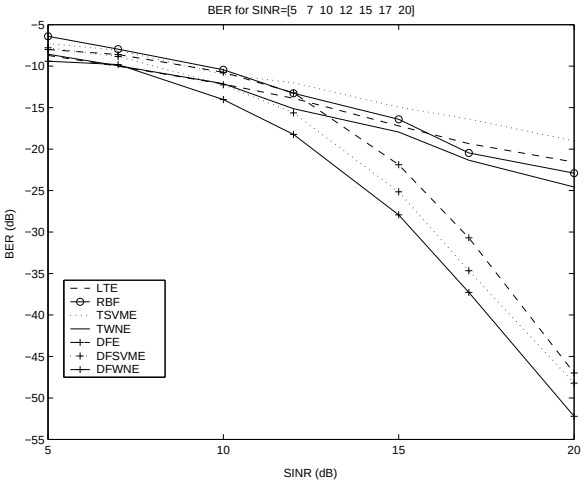


Figure 4: Comparison between the DFSVME, DFWNE, LTE, DFE, and RBF in case 2

C \ d	2	3	4	5
1	-16.1	-16.0	-14.5	-13.6
2	-15.9	-16.1	-15.0	-13.4
5	-16.4	-16.4	-14.0	-14.6
10	-16.2	-15.1	-13.8	-13.9
50	-14.0	-15.6	-13.9	-14.3
∞	-16.3	-15.3	-14.1	-14.2

Table 6: BER results with different values of C and d for the SVM-based equalizer (feedback case)

resented by $H_0(z) = 0.3482 + 0.8704z^{-1} + 0.3482z^{-}2$ and $H_1(z) = \lambda(0.6 + 0.8z^{-1})$, is used to compare the performance of the equalizers. The Transversal Support Vector Machine(SVM) Based Equalizer (TSVME), the Decision Feedback SVM-based Equalizer (DFSVME), the Transversal Wavelet Network-based Equalizer (TWNE), the Decision Feedback Wavelet Network-Based Equalizer(DFWNE) are all compared with the conventional Linear Transversal Equalizer(LTE), Decision Feedback Equalizer(DFE), and the Radial Basis Function Equalizer (RBF) for the same system. For a fair comparison, all the transversal equalizers employed were chosen to be of order 4 and the equalization delay was set to ($\tau = 1$). The decision feedback equalizers were simulated with a feedforward and feedback order of 2 and 2, respectively, and an identical delay $\tau = 1$. All experimental results come from the averaging of ten independents runs over a thousand hundred bipolar test samples.

The two Wavelet Network-based equalizers were trained on-line using an incremental gradient-descent algorithm. The training was stopped when the error had reached a steady state. In the case of the $TSVME$ and the $DFSVME$, all the training patterns and targets were used to solve the quadratic problem. Then the training patterns with zero-valued Lagrange multipliers were pruned, and only those left were taken in account during the testing phase.

The best configuration for the considered equalizers were determined beforehand using a trail and error basis (namely the Nw (hidden-layer size) and choice of the transfer function for the Wavenet, and the value of C and the degree of the polynomial kernel for the case of the SVM). These trial results were computed with a $SNR = 15dB$ and a $SIR = 15dB$ (which leads to a global $SINR \simeq 11.99dB$).

As shown in Tables 3 and 4, the Wavenet-based equalizer offers better results with a sigmoidal transfer function in all cases. Consequently, only networks using a sigmoidal transfer function where further evaluated. The effect of the parameter Nw was tested between the range 2 and 8, since higher values would certainly allow the network to over-fit the data, and lead to too high computational complexity. One can observe two falls at $Nw = 3$ and $Nw = 7$, for both the Transversal (Feedforward) and Feedback case, with very close BER values. In the first instance, a value of $Nw = 3$ was chosen, as the smaller the size of the hidden layer, the less free parameters the network would have. But further simulation results showed that a hidden layer of size $Nw = 3$ was not complex enough, and offered poor performance in high $SINR$ case. Consequently, a value of $Nw = 4$ was finally used for later experiments.

Considering the SVM-based equalizer (see Tables 5 and 6), the greater the value of d, worst the results. A value of $d = 3$ was chosen, as it offers very similar results to $d = 2$, and during experiments with low $SINR$'s, a more complex decision boundary would be necessary. The value of C seems to be irrelevant, since there are very little differences between the results in the same column (in Table 5 and 6). A value of $C = 5$ was thus chosen, as it offered good general results, either in the transversal or the decision-feedback case.

During the comparison with others equalizers, two scenarios where simulated as described below:

1. a value of $\lambda = 0.0631$ was chosen to provide a constant $SIR = 24dB$, and the noise power was varied to produce different $SINR$'s.

2. the noise power was fixed to $\sigma_n^2 = 0.00398$ giving rise to constant $SNR = 24dB$. Then, the interfering signal power was changed by choosing different values of λ.

The four proposed equalizers were compared to other common equalizers: the LTE, the DFE, and the RBF. Results are shown in figure 3 for case 1, and in figure 4 for the second case. As can be seen, the DFWNE gives the best BER performance, followed by the DFSVME. Surprisingly, The TSVME provides the worst performance at all, which need further investigation.

5 Conclusion

A realistic co-channel system was used as a case study to demonstrate the equalization capability of the four novel architectures, namely the SVM-based equalizers and the Wavelet Network-based equalizers. The results have shown better BER performance characteristics for the two of them incorporating decision feedback equalization, comparatively to the common LTE, DFE, and RBF equalizers.

The results in this study have considered single co-channel systems, but they can be readily extended to the multi-co-channel case.

For future work, the performance of the proposed equalizers need to be compared with the optimal Bayesian and MLVA based approaches for combating the co-channel interference problem in mobile cellular networks.

References

[1] Amir Hussain, John J. Soraghan, and Tariq S. Durrani, "Adaptive Functional-Link Neural-Network-Based Non-Linear Equalizers for Overcoming Co-Channel Interference", IEEE international workshop on signal processing methods in multipath environment, pp.105-113 (1995)

[2] Y. Oussar, I. Rivals, L. Personnaz, G. Dreyfus "Training Wavelet Networks for Nonlinear Dynamic Input-Output Modeling", Neurocomputing, in press. (1998)

[3] A. Graps, "An introduction to Wavelets" IEEE Computational Science and Engineering, vol. 2, num. 2 (Summer 1995)

[4] Qinghua Zhang, "Using wavelet network in nonparametric estimation", IRISA, publication interne N.833 (June 1994)

[5] A. Ypam, R. P.W. Duin, "Using the Wavenet for function approximation", Pattern Recognition Group, Faculty of Applied Physics, Delft University of Technology Lorentzweg 1, 2628 CJ, Delft, The Netherlands.

[6] C. Cortes and V. Vapnik, "Support-Vector Networks", Machine Learning, Vol. 20, pp.273-297 (1995)

[7] Steve R. Gunn, "Support Vector Machine for classification and regression", Technical report, ISIS Research Group, Department of Electronics and Computer Science, University of Southampton, UK (May 1998)

[8] M. O. Stitson, J. A. E. Wetson, A. Gammerman, V. Vovk, and V. Vapnik, "Theory of Support Vector Machine", Technical Report, Department of Computer Science, University of London (December 1996)

[9] Christopher J.C. BURGES, "A tutorial on Support Vector Machine for Pattern Recognition", Data Mining and Knowledge Dicovery 2, 121-167 (1998)

[10] S. Chen, S. Gunn, and C.J. Harris, "Decision Feedback Equaliser Design Using Support Vector Machine", IEE Proceedings Vision, Image and Signal Processing 147(3): 213-219 (2000)

[11] D.J. Sebald "Support Vector Machine Techniques for Nonlinear equalization" IEEE transactions on signal processing, Vol. 48, No. 11, pp. 3217-3226 (November 2000)

Complementary Beamforming

Vahid Tarokh
Harvard University
33 Oxford Street Room MD-347
Cambridge, MA 02139

Yang-Seok Choi and Siavash M. Alamouti
Vivato Research and Development
12610 E. Mirabeau Parkway Suite 900
Spokane, WA 99216

Abstract

We introduce *Complementary Beamforming* (CBF) for wireless communications which tailors beamforming in order to provide an overall superior network performance. An application of complementary beamforming is in smart antennas enhancements to the IEEE 802.11 systems. In these systems, beamforming may be employed in order to increase the range/capacity. When a conventional beamformer is employed to increase the transmitted power to some specific directions, the radiated power to other directions is reduced. This means that some other users of the system may experience lower received signal levels. During a busy period of the channel, these users may wrongly determine that the channel is idle and start transmitting packets. This may cause unnecessary transmissions, subsequent back-offs, increased network latency and interference. Furthermore, the aforementioned undesired packet transmission has an energy penalty which adversely effects the battery life of the remote devices. This "*Hidden Beam Problem*" causes more severe degradation if the system is more heavily loaded which will be most likely the case both in hot spots and also whenever the system range is increased. In this work, we apply complementary beamforming and construct techniques that are designed to significantly reduce the probability of the aforementioned collisions, back-offs and re-transmissions. We analyze the proposed scheme for both the intended and silent users and prove that, when compared to conventional methods, for a meager incurred power loss for the intended users, the effects of the hidden beam problem can be significantly reduced. Moreover, we will show that the complementary beamforming described in this paper are approximately twice as complex as conventional beamforming techniques. Finally, we will illustrate the proposed technique by some examples.

1 Introduction

Theoretically, using directional signals (beams), it is possible to increase the range or capacity of any wireless link without changing the air-interface or even increasing the transmit power. This is done by focusing the energy in the direction of desired receivers (nodes) using antenna arrays and beamforming networks. As a result, there has been great interest in using directional signals

to improve the performance of wireless systems. In fact, most commercial mobile radio networks employ sectorization which is a simple form of directional signal transmission.

Directional signals can make a big difference in the efficiency and performance of wireless networks. Only in broadcast systems, the transmitted data is targeted to more than one node. In most wireless networks, data-bearing signals are intended to travel between two nodes; typically a central node and a portable or mobile node. In such networks, it is inefficient to transmit a signal in all directions. The energy in all directions except in the direction of the desired node is wasted. This waste in energy limits the performance of most commercial wireless networks. For instance, the 54 Mbit/sec mode of the 802.11a standard has a range of only a few meters in typical deployment scenarios with omnidirectional antennas. Using directional signals, the range (or signal quality) can be significantly increased. Moreover, directional signals reduce the overall interference in the network.

Unfortunately, however, most commercial wireless random access networks including 802.11 rely on channel sensing for their access mechanism. For instance, the 802.11 standards use Carrier Sense Multiple Access (CSMA) which is a listen-before-talk scheme. Beamforming has the side-effect of hiding the transmitted signal from some nodes in the network. We have called this phenomenon the *hidden beam problem*. Typically, a given node listens for energy from other nodes in the network. If it cannot detect the presence of other transmissions, it attempts to gain access to the medium. If not addressed, the hidden beam problem can cause unnecessary transmissions and back-offs which negatively impact the performance of the network.

In other words, in channel sensing networks, every transmission carries some useful information for all the nodes in the network. The transmitted signal carries data to its target node, and also informs the rest of the nodes in the network not to transmit. Therefore, beamforming can potentially destroy some valuable information intended for some nodes.

Fortunately, in practice, directional signals are not "pencil beams". They generally have a main lobe whose width is constrained by the antenna structure, and sidelobes whose levels vary in different directions. Nevertheless, these beams may have deep nulls in some directions. In these directions, the network will suffer from the hidden beam problem. For these networks to operate efficiently, we would like the desired (active) nodes to have sufficient signal to interference and noise ratio (SINR) to decode the data and at the same time for the rest of the nodes (passive nodes) to receive sufficient energy to refrain from transmission.

In an ideal scenario, we would like to establish simultaneous spatial links to active nodes using Space Division Multiple Access (SDMA) and a broadcast link to the rest of the nodes in the network (passive nodes). The broadcast link would carry control information for all the passive nodes in

the network. The concept of transmitting data in one or more simultaneous beams combined with a broadcast signal (whether a non-data-bearing energy signal or an actual data-bearing signal) to all other directions (herein referred to by the *Complementary Region*) is what we have termed complementary beamforming.

In this paper, we report a technique where, using SDMA, multiple beams are transmitted to active nodes while, using complementary beamforming, some energy is transmitted to all the passive nodes not covered by the main beams. In [1], we will present a simple scheme for applications where SDMA is not required, where we control the sidelobe levels on a single transmit beam to ensure a minimum level of energy transmitted in all directions. One obvious application of these techniques is to reduce the effect of the hidden beam problem for networks with channel sensing. Since a given receiver's energy detection threshold is usually lower than its decoding threshold, it is possible to focus the signal towards one or more active nodes (through main beams) and yet ensure minimum transmit power (through complementary beams) towards other nodes in the network (within the complementary regions). This would increase the probability that the signal is detected by all the nodes. In some other applications, it may be desirable for hidden beam nodes to decode some of the transmit data. To this end, we have devised a technique called *Complementary Superposition Beamforming* which will be disclosed in [3].

An immediate application of these techniques is to smart antenna enhancements to IEEE 802.11 wireless LANs. These standards support a maximum throughput of 54 Mbps. However, propagation studies and measurements, such as those reported in [2, 5] have shown that the range of typical Wi-Fi equipments at these data rates are at best limited to a few meters. At present, a "Wi-Fi revolution" is taking place and it is expected that the widespread deployment of Wi-Fi will change the entire wireless landscape in few years. Abundance of data hungry users in *hot spots* will motivate wireless LAN providers to seek WLAN devices with increased throughputs and ranges. This has created significant interest and activities in the wireless industry focusing on techniques to increase the range and capacity of Wi-Fi networks. As mentioned before, an enhancement that seems to provide an appealing solution is the use of antenna arrays at access points (AP) in conjunction with beamforming since such a solution is appealing as it is transparent to receivers and does not force any changes to current standards. It has been verified that complementary beamforming can significantly enhance the performance of such a system [6].

The outline of this paper is given next. In Section 2, we will establish the notation, and discuss the *hidden beam problem* in CSMA Systems. In Section 3, we give an approach to complementary beamforming. In Section 4, we analyze the performance of the proposed scheme for both the intended and silent users. It will be proved that, when compared to conventional methods, for a

meager incurred power loss for the intended users, the effects of the hidden beam problem caused can be significantly reduced. Moreover, we will show that the complementary beamforming described in this paper are approximately twice as complex as conventional beamforming techniques. In Section 5, we will illustrate the proposed technique by some examples. Finally, some conclusions and directions for future research will be provided in Section 6.

2 The Hidden Beam Problem

To illustrate the situation, let us consider a scenario when a system employs m transmit antennas and the transmitter simultaneously transmits to k users. In this work, we assume that $k \leq m$.

A conventional beamformer seeks to increase the power pointed to the k desired users. Consider a scenario where there are $m = 2$ transmit antennas and $k = 1$ intended users. Let the channel matrix to the desired user be given by (α, β). A conventional beamformer then induces weigths $w_1 = \frac{\bar{\alpha}}{\sqrt{|\alpha|^2+|\beta|^2}}$ and $w_2 = \frac{\bar{\beta}}{\sqrt{|\alpha|^2+|\beta|^2}}$ at the transmitter, where $\bar{\alpha}$ and $\bar{\beta}$ are the conjugates of α and β respectively.

If c_1 is the intended transmit signal at time 1 for user 1, then $w_1 c_1$ and $w_2 c_1$ are transmitted signals from antennas 1 and 2 respectively. The intended user receives the signal

$$r_1 = w_1 \alpha c_1 + w_2 \beta c_1 + n_1 = \sqrt{|\alpha|^2 + |\beta|^2} c_1 + n_1, \tag{1}$$

where n_1 is the noise. It is immediately observed that the signal to noise power ratio of the desired user improves by a factor of $10 \log_{10}(|\alpha|^2 + |\beta|^2)$ dB.

The above power improvement does not come for free. Let an unintended user have channel matrix $(-\bar{\beta}, \bar{\alpha})$. Then the signal at this unintended user is given by

$$y_1 = -\bar{\beta} w_1 c_1 + \bar{\alpha} w_2 c_1 + \eta_1 = \eta_1, \tag{2}$$

where η_1 is the noise vector and the unintended user receives no signal.

We observe that there is no part of the transmitted signal present at this unintended user's receiver. This by itself may not seem to pose a serious problem, since after all the transmission was not intended for this user. But it turns out that it can cause a problem in beamforming enhancements to CSMA based systems such as those based on the IEEE 802.11 WLAN standard. In these systems, all users and the access point share the same channel for both uplink and downlink transmissions. Each user senses the channel and only transmits packets if it determines that the channel is not busy. The unintended receiver who does not receive a strong signal component may wrongly determine that the channel is idle and start transmitting packets. This may cause unnecessary

transmissions, subsequent back-offs, increased network latency and interference. Furthermore, the aforementioned undesired packet transmission has an energy penalty which adversely effects the battery life of the remote devices. This "*Hidden Beam Problem*" causes more severe degradation if the system is more heavily loaded which will be most likely the case both in hot spots and also whenever the system range is increased. We refer to this problem as *the hidden beam problem.*

3 Complementary Beamforming

In this paper, we consider the hidden beam problem and propose an elegant solution which we refer to as *complementary beamforming.* The main intuition behind complementary beamforming is that much less power is needed for an unintended user to correctly detect a busy period than that required for correct detection of the transmitted packet. The fact that detecting channel activity is much simpler than decoding the received word is well understood in communications [7]. This is because an error in detection of channel activity happens when a transmitted codeword is confused with the all zero signal. In contrast a decoding error is made when the transmitted signal is confused with other codewords. We can also arrive at the above conclusion using tools of information theory [4]. At code rates above the channel capacity, Shannon has proved that the block decoding error probability asymptotically tends to one and that the bit error rate is bounded below by a positive number. However, even at transmission rates above capacity, it is easy to observe that the probability of channel activity detection error asymptotically goes to zero as the block length goes to infinity.

The above observation is built in the detection criteria for channel activity in IEEE 802.11 WLAN standards. Each device listens to the channel during some time window and compares the energy collected in this window to a value called the CCA (Clear Channel Assessment) threshold. Activity is detected only if the collected energy is greater than the CCA threshold.

From the above, we observe that any IEEE 802.11 device requires much less receive power to correctly determine channel activity than to decode the transmitted signals. This motivates our solution to the hidden beam problem. We will seek to construct a beam pattern which directs most of the transmitted power to the intended recipients while directing a small fraction of the total power to unintended users. Once such a beam pattern is designed, the unintended users will all sense the transmission to the desired users with high probability and will keep silent during a busy downlink period. This in turn reduces the packet collision probability.

Next we construct such a beam pattern. To this end, we first introduce some notations.

Notation:

- δ_j denotes a k-dimensional column vector with j-th component equal to 1 and other components equal to zero.
- For any vector X, we let X^T and X^H respectively denote the transpose and Hermitian of X.
- For any matrix D, we let W_D denote the vector space spanned by the columns of D.
- Let the channel from transmit antenna l to the intended user j be given by $\alpha_{l,j}$.
- Let A_j denote the column vector $(\alpha_{1,j}, \alpha_{2,j}, \cdots, \alpha_{m,j})^T$. We refer to the vector A_j as the *spatial signature of user j*.
- Let A denote the matrix whose j-th column is A_j.
- Let $R^t = (r_1^t, r_2^t, \cdots, r_k^t)$ and $X^t = (x_1^t, x_2^t, \cdots, x_m^t)$ respectively denote the vector of received signals at intended users $j = 1, 2, \cdots, k$ and the vector of signals transmitted from antennas $1, 2, \cdots, m$ at time t.
- Let $C^t = (c_1^t, c_2^t, \cdots, c_k^t)$, where c_j^t is the signal intended to the $j = 1, 2, \cdots, k$ desired user at time t.
- For any square matrix A, let $Tr(A)$ denotes the trace (sum of diagonal elements of A).
- Let $N^t = (n_1^t, n_2^t, \cdots, n_m^t)$ be the noise vector components at time t at the intended users.

Then it is well-known that

$$R^t = X^t A + N^t, \tag{3}$$

In most cases, these noise components are assumed to be i.i.d. Gaussian with variance σ^2 per complex dimension. We make *absolutely no assumptions* on the statistics of the matrix A.

It will be assumed that c_j^t, $j = 1, 2, \cdots, k$, $t = 1, 2, \cdots, L$ are elements of a signal constellation with average signal $E[c_j^t] = 0$. We will also assume that the elements of the signal constellation are normalized so that their average power is $E[|c_j^t|^2] = 1$.

In general $X^t = C^t \mathcal{B}$ where $\mathcal{B}$ is referred to as the *beamforming matrix*. The choice of $\mathcal{B}$ depends on the beamforming strategy and many approaches for the selection of $\mathcal{B}$ are suggested in the literature. Assuming that the matrix A is known at the transmitter and the existence of $(A^H A)^{-1}$, for a zero-forcing beamformer

$$\mathcal{B} = \frac{(A^H A)^{-1} A^H}{\sqrt{Tr((A^H A)^{-1})}}$$

and

$$X^t = \frac{C^t (A^H A)^{-1} A^H}{\sqrt{Tr((A^H A)^{-1})}}. \tag{4}$$

Another commonly used beamforming matrix is given by

$$\mathcal{B} = \frac{(A^H A + \frac{1}{\text{SNR}} I)^{-1} A^H}{\sqrt{Tr((A^H A + \frac{1}{\text{SNR}} I)^{-2} A^H A)}}, \tag{5}$$

where $SNR = \frac{1}{\sigma^2}$. Other choices of beamforming matrices are also possible.

Under the above assumptions the total transmit power is easily computed to be 1. For simplicity, we present our technique for the zero-forcing beamformer here. Nonetheless, we note that the method that we present here generalizes to other cases as well. In the following, we will sometimes describe this generalization. Moreover in our presentation, we also assume that the spatial signature matrix A is constant during the transmission of a packet and varies from one packet to another.

Assuming a zero-forcing beamformer, the received signal at the receiver is given by

$$R^t = \frac{C^t}{\sqrt{Tr((A^H A)^{-1})}} + N^t,$$

and we observe that each intended user $j = 1, 2, \cdots, k$ receives a noisy version of its intended signal scaled by a factor $Tr((A^H A)^{-1})$.

It is readily observed that if an unintended user has spatial signature $B = (b_1, b_2, \cdots, b_m)^T$ orthogonal to all the rows of A, then it receives the signal

$$y^t = X^t B + \eta^t = C^t (A^H A)^{-1} A^H B / \sqrt{Tr((A^H A)^{-1})} + \eta^t = \eta^t,$$

at time t, where η^t is Gaussian noise. This means that such a user does not receive any signal components. As mentioned above, such an unintended user can confuse a busy downlink period with a silent period and transmit packets during a busy period. This can cause unwanted collisions and reduce the efficiency of the system. Thus we arrive at a simple albeit important conclusion that, *whenever a $k \times m$ beamforming matrix is fixed during transmission of a packet, then any unintended user that has spatial signature in the orthogonal complement of the subspace generated by the rows of the beamforming matrix receives no signal components.* In time domain, this motivates the use of different beamforming matrices at different instances of time during the transmission of downlink packets, so that the effects of the hidden beam problem can be reduced.

3.1 The Proposed Scheme:

To this end, we observe that the subspace W_A is a k-dimensional subspace of the complex m-dimensional complex space and has an orthogonal complement $W_A^{\perp}$ of dimension $m - k$. Let

$U_0, U_1, \cdots, U_{m-k-1}$ form an orthonormal basis for $W_A^{\perp}$. In other words, $U_0, U_1, \cdots, U_{m-k-1}$ are mutually orthogonal m-dimensional column vectors of length one in $W_A^{\perp}$. Clearly, $U_j^H A_i = 0$ for $0 \leq j \leq m-k-1$ and $1 \leq i \leq k$.

First, the transmitter constructs matrices $Z_1, Z_2, \cdots, Z_L$, where L is the the length of downlink transmission period, such that these matrices satisfy the following properties.

- **A:** For all $1 \leq i \leq L$, the matrix Z_i is a $k \times m$ matrix whose rows are in the set

$$\{0, \pm U_0^H, \pm U_1^H, \cdots, \pm U_{m-k-1}^H\},$$

- **B:** If L is even, then $Z_2 = -Z_1$, $Z_4 = -Z_3, \cdots, Z_L = -Z_{L-1}$,

- **C:** If L is odd, then $Z_2 = -Z_1$, $Z_4 = -Z_3, \cdots, Z_{L-1} = -Z_{L-2}$, $Z_L = 0$, and

- **D:** Each element

$$+U_0^H, -U_0^H, +U_1^H, -U_1^H, \cdots, +U_{m-k-1}^H, -U_{m-k-1}^H$$

  appears p times in the the list of Lk rows of $Z_1, Z_2, \cdots, Z_L$ for some positive integer p. If this cannot be exactly satisfied, we try to have the number of these appearances as large and as close as possible to each other. Clearly, for $p = \lfloor k(L-1)/2(m-k) \rfloor$, it is possible to have p occurrences of each of these vectors and r occurrences of the zero vector, where $r = LK - 2p(m-k)$.

From Property **D**, it is immediately observed that

$$kL - 2m - k \leq 2p(m-k) \leq 2\lfloor \frac{L}{2} \rfloor k. \tag{6}$$

Because $p \geq 1$, from the above inequality, we observe that for $L < \frac{2(m-k)}{k}$, Property **D** cannot be exactly satisfied. Thus, for extremely short packets, we cannot always provide a perfectly balanced appearance of $+U_0^H, -U_0^H, +U_1^H, -U_1^H, \cdots, +U_{m-k-1}^H, -U_{m-k-1}^H$.

In practice, there are a number of easy ways to implement construction of matrices of $Z_1, Z_2, \cdots, Z_L$ that approximately or exactly satisfy Property (**D**). An straightforward construction is:

- For instance for even L, if we let $p = \lfloor Lk/2(m-k) \rfloor$, then we can assign U_j^H, $j = 0, 1, \cdots, (m-k-1)$ in sequential manner to the first p available columns of $Z_1, Z_3, \cdots$ and replace the remaining columns with zeroes. We then let $Z_2 = -Z_1$, $Z_4 = -Z_3, \cdots, Z_L = -Z_{L-1}$

- For odd L, we construct the $L-1$ matrices $Z_1, Z_2, \cdots, Z_{L-1}$ as above and let $Z_L = 0$.

In both cases, we can guarantee that

$$\frac{Lk}{2(m-k)} \geq p \geq \lfloor \frac{k(L-1)}{2(m-k)} \rfloor \tag{7}$$

which gives Inequality (6).

It can be easily seen that the matrices $Z_1, Z_2, \cdots, Z_L$ given by the above construction satisfy the above Properties. Other constructions are also possible.

Once $Z_1, Z_2, \cdots, Z_L$ are constructed, at each time t, the transmitter chooses the beamforming matrix

$$S^t = [((A^H A)^{-1} A^H / \sqrt{Tr((A^H A)^{-1})} + \frac{1}{\sqrt{k}} \epsilon Z_t], \tag{8}$$

where $\epsilon \geq 0$ is a fixed positive number. The choice of $\epsilon \geq 0$ governs the trade-off between the power pointed to the intended users and that pointed to unintended users. By increasing the power pointed to intended users, the intended users enjoy better channels, while by pointing more power to unintended users, better channel activity detection during the busy periods can be achieved. This trade-off will be analyzed in the next section and criteria for the choice of $\epsilon \geq 0$ will be determined. For $\epsilon = 0$, we recover the conventional beamforming. Thus complementary beamforming generalizes and includes conventional beamforming as a special case.

We note that in the proposed scheme the beamforming matrix varies from one time to another. This guarantees that a small fraction of power is pointed to every direction of the space and that the unintended receivers can determine channel activity periods with higher probabilities.

4 Analysis of Complementary Beamforming

We analyze complementary beamforming scheme both for the intended and unintended receivers.

4.1 The Power Penalty for The Intended Users:

The addition of the term $\frac{1}{\sqrt{k}} \epsilon Z_i$ to the matrix $(A^H A)^{-1} A^H / \sqrt{Tr((A^H A)^{-1})}$ increases the transmit power. To compute the penalty, we use the orthogonality of $U_0, U_1, \cdots, U_{m-k-1}$ and the columns of A to conclude that $Z_t A = 0$ for all $t = 1, 2, \cdots, L$. Thus, we compute the receive word for intended users to be

$$R^t = C^t S^t A + N^t = \frac{C^t}{\sqrt{Tr((A^H A)^{-1})}} + N^t,$$

which is the same as the conventional beamforming. In contrast, in the case of complementary beamforming, we use the matrix equality

$$Tr\left[(Y+W)(Y+W)^H\right] + Tr\left[(Y-W)(Y-W)^H\right] = 2Tr(YY^H) + 2Tr(WW^H),$$

and Properties **B** and **D** to compute the average transmitted power

$$\frac{\sum_{t=1}^{L} Tr(S_t S_t^H)}{L} = 1 + \frac{\sum_{t=1}^{L} Tr(Z_t Z_t^H)}{Lk}|\epsilon|^2.$$

From Property **D**, we have

$$\sum_{t=1}^{L} Tr(Z_t Z_t^H) = 2p(m-k),$$

thus

$$\frac{\sum_{t=1}^{L} Tr(S_t S_t^H)}{L} = 1 + \frac{2p(m-k)}{Lk}|\epsilon|^2.$$

We can now prove the following Theorem.

Theorem 1 *The intended users in complementary beamforming when compared to the conventional method suffer a loss of at most* $10\log_{10}(1+|\epsilon|^2)$.

Proof: This follows from the above and from inequality (6). □

4.2 Analysis of The Power delivered to Silent Users:

Let $B = (b_1, b_2, \cdots, b_t)^T$ denote the channel of an arbitrary unintended user. We will next study the power received by this unintended user under complementary beamforming. To this end, we recognize that the columns of matrix A and the vectors $U_0, U_1, \cdots, U_{m-k-1}$ span the complex m-dimensional space. Thus we can write

$$B = e_1 A_1 + \cdots + e_k A_k + d_0 U_0 + \cdots + d_{m-k-1} U_{m-k-1}, \tag{9}$$

for some constants $e_1, e_2, \cdots, e_k$ and $d_1, d_2, \cdots, d_{m-k}$. Computing $B^H B$, we arrive at

$$\sum_{j=1}^{m} |b_i|^2 = (e_1^H, e_2^H, \cdots, e_k^H) A^H A (e_1^H, e_2^H, \cdots, e_k^H)^H + \sum_{j=0}^{m-k-1} |d_j|^2. \tag{10}$$

At time t, the unintended receiver now receives

$$y^t = X^t B + \eta^t = C^t S^t B + \eta^t.$$

By replacing for S^t and B from Equations (8) and (9) and observing that

$$(A^H A)^{-1} A^H A_j = \delta_j,$$

$$A^H U_i = 0$$

$$Z_t A_i = 0,$$

we arrive at the conclusion that

$$S^t B = \frac{(e_1^H, e_2^H, \cdots, e_k^H)^H}{\sqrt{Tr((A^H A)^{-1})}} + \frac{\epsilon}{\sqrt{k}} \sum_{j=0}^{m-k-1} d_j Z_t U_j. \tag{11}$$

We next compute the average expected receive signal power

$$P_{av} = \frac{\sum_{t=1}^{L} E[|y^t|^2]}{L} = \frac{\sum_{t=1}^{L} Tr(S^t B B^H (S^t)^H)}{L}. \tag{12}$$

However, since $Z_{2l} = -Z_{2l-1}$ for $l = 1, 2, \cdots, \lfloor \frac{L}{2} \rfloor$ is assumed, we can use Equation (11) and simple manipulations to arrive at

$$Tr(S^{2l} B B^H (S^{2l})^H + S^{2l-1} B B^H (S^{2l-1})^H) = \frac{2\sum_{j=1}^{k} |e_j|^2}{Tr((A^H A)^{-1})} +$$
$$\frac{|\epsilon|^2}{k} \sum_{j=0}^{m-k-1} |d_j|^2 \left[Tr(Z_{2l-1} U_j U_j^H Z_{2l-1}^H) + Tr(Z_{2l} U_j U_j^H Z_{2l}^H) \right]$$

Using the above and after simple manipulations, we arrive at

$$P_{av} = K(L) \frac{\sum_{j=1}^{k} |e_i|^2}{Tr((A^H A)^{-1})} + \frac{|\epsilon|^2}{kL} \sum_{j=0}^{m-k-1} |d_j|^2 \sum_{t=1}^{L} Tr(Z_t U_j U_j^H Z_t^H),$$

where $K(L) = 2\lfloor L/2 \rfloor / L$. The sum $\sum_{t=1}^{L} Tr(Z_t U_j U_j^H Z_t^H)$ is exactly equal to the number of times that $\pm U_j$ appears in the list of the rows of $Z_1, Z_2, \cdots, Z_L$. By Property **D** this amounts to $2p$. Thus

$$P_{av} = K(L) \frac{\sum_{j=1}^{k} |e_i|^2}{Tr((A^H A)^{-1})} + |\epsilon|^2 \frac{2p}{kL} \sum_{j=0}^{m-k-1} |d_j|^2. \tag{13}$$

We now proceed to lower bound P_{av}. To this end, we prove the following theorem.

Theorem 2 *Let $\lambda_{\min}(A^H A)$ and $\lambda_{\max}(A^H A)$ respectively denote the minimum and maximum eigenvalues of $A^H A$. Then provided that*

$$|\epsilon|^2 \leq \frac{(m-k)}{k} \frac{\lambda_{\min}(A^H A)}{\lambda_{\max}(A^H A)}, \tag{14}$$

$$p \geq \frac{m}{k} - 0.5, \tag{15}$$

complementary beamforming guarantees a fraction $|\epsilon|^2 \frac{\sum_{j=1}^{m} |b_j|^2}{m}$ of the transmitted power to an unintended receiver whose spatial signature is $B = (b_1, b_2, \cdots, b_m)$.

Proof: Let an unintended user with spatial signature given by $B = (b_1, b_2, \cdots, b_m)$ be given. Suppose that the inequality (14) holds. From Equations (10) and (13), we observe that

$$P_{av} = |\epsilon|^2 \frac{2p}{kL} \sum_{i=1}^{m} |b_i|^2 - (e_1^H, e_2^H, \cdots, e_k^H) G (e_1^H, e_2^H, \cdots, e_k^H)^H,$$

where

$$G = \left[(|\epsilon|^2 \frac{2p}{kL} A^H A - \frac{K(L) I}{Tr((A^H A)^{-1})} \right],$$

and I is the identity matrix. The matrix G is Hermitian, thus we conclude from the above that

$$P_{av} \geq |\epsilon|^2 \frac{2p}{kL} \sum_{i=1}^{m} |b_i|^2 - \lambda_{\max}(G) \sum_{j=1}^{k} |e_k|^2 \tag{16}$$

where $\lambda_{\max}(G)$ is the maximum eigenvalue of G. Clearly

$$\lambda_{\max}(G) = |\epsilon|^2 \frac{2p}{kL} \lambda_{\max}(A^H A) - \frac{K(L)}{Tr((A^H A)^{-1})}.$$

Next, we prove that $\lambda_{\max}(G) \leq 0$. Clearly

$$Tr((A^H A)^{-1}) \leq \frac{k}{\lambda_{\min}(A^H A)},$$

thus using Condition (14)

$$\frac{1}{Tr((A^H A)^{-1})} \geq \frac{\lambda_{\min}(A^H A)}{k} \geq \frac{|\epsilon|^2 \lambda_{\max}(A^H A)}{m-k},$$

which gives

$$|\epsilon|^2 \frac{2p}{kL} \lambda_{\max}(A^H A) \leq \frac{2p(m-k)}{kL} \frac{1}{Tr((A^H A)^{-1})} \leq \frac{K(L)}{Tr((A^H A)^{-1})},$$

where we used inequality (6). We conclude from the above that $\lambda_{\max}(G) \leq 0$. Using Equation (16), this implies that

$$P_{av} \geq |\epsilon|^2 \frac{2p}{kL} \sum_{i=1}^{m} |b_i|^2 \geq |\epsilon|^2 \frac{2pm}{kL} \frac{\sum_{i=1}^{m} |b_i|^2}{m}. \tag{17}$$

Using the inequality (6) and the condition $p \geq \frac{m}{k} - 0.5$, we can now conclude that $P_{av} \geq |\epsilon|^2 \frac{\sum_{i=1}^{m} |b_i|^2}{m}$.

□

Discussion: Intuitively, the condition $p \geq \frac{m}{k} - 0.5$ means that the transmitted packets must not be too short. However, it may not seem intuitive to the reader that the Condition (14) on ϵ contains terms of the form $\lambda_{\min}(A^H A)/\lambda_{\max}(A^H A)$. We argue that this condition is intuitively appealing. To this end, consider the case that the ratio $\lambda_{\min}(A^H A)/\lambda_{\max}(A^H A)$ is small. Then the matrix $A^H A$ is close to being singular. This means that even the intended users, do not receive significant signal powers. In fact, practical beamforming schemes, when scheduling transmission to intended users always assure that the ratio $\lambda_{\min}(A^H A)/\lambda_{\max}(A^H A)$ is not close to zero and in most cases even larger than a pre-specified threshold. In practice, a ratio $\lambda_{\min}(A^H A)/\lambda_{\max}(A^H A) \geq \frac{1}{3}$ is generally an acceptable assumption. In the case of system with $k = 4$, $m = 16$. Thus, provided that scheduling algorithm can guarantee that $\lambda_{\min}(A^H A)/\lambda_{\max}(A^H A) \geq \frac{1}{30}$, the above complementary beamforming scheme could be used to provide any fraction $|\epsilon|^2 \leq 0.1$ of the transmitted power to unintended users.

5 Examples

Example I: We consider the case when there are $m = 2$ transmit antennas and $k = 1$ intended receivers. Assuming that the channel to the intended user is given by $A = (\alpha, \beta)^T$, we observe that $\lambda_{\min}(A^H A)/\lambda_{\max}(A^H A) = 1$ and as long as $|\epsilon|^2 \leq 1$, by the above theorem a fraction $|\epsilon|^2$ of the transmitted power is pointed to unintended users at the expense of a loss of at most $10 \log_{10}(1+|\epsilon|^2)$ to the intended user. With $\epsilon = 0.1$, we observe that a power of 20 dB below transmit power can be guaranteed to any unintended users so that they can detect channel activity, while the power penalty for the intended user is only 0.044 dB.

The beamforming matrices S_1 and S_2 in this case are given by

$$S_1 = \frac{1}{\sqrt{|\alpha|^2 + |\beta^2|}}(\bar{\alpha} - \epsilon\beta, \bar{\beta} + \epsilon\alpha),$$
$$S_2 = \frac{1}{\sqrt{|\alpha|^2 + |\beta^2|}}(\bar{\alpha} + \epsilon\beta, \bar{\beta} - \epsilon\alpha),$$

with $S_{2l-1} = S_1$ and $S_{2l} = S_2$ for $l = 1, 2, \cdots, \lfloor \frac{L}{2} \rfloor$ when the transmission period is of length L with $S_L = \frac{1}{\sqrt{|\alpha|^2+|\beta^2|}}(\bar{\alpha}, \bar{\beta})$ when L is odd.

In Figures 1,2,3,4,5, and 6, we have compared CBF with conventional beamforming using the above scheme. In Figures 1,2,3 and 4, the distance d between the transmit antennas is set to be half of the wavelength λ. In Figures 1 and 2, the value of $\epsilon = 0.3$ and in Figures 3 and 4, $\epsilon = 0.1$. Beam patterns in time are illustrated and the average power value (CBF) is compared to the conventional case. It is immediately seen from the Figures, that with a meager power penalty to the intended user, the hidden beam problem is completely eliminated. Similarly, Figures 5 and 6 correspond to $\epsilon = 0.3$ and $d = \lambda$. Figures 7,8 provide the beam patterns for complementary beamforming for higher number of antennas.

Example II: In this example, we will give a construction of the complementary beamforming matrices for the case that there are $m = 16$ transmit antennas and $k = 4$ intended receivers. We also compare the complexity of complementary beamforming to that of the conventional beamforming.

The channel matrix A is a 16×4 matrix. The columns of this matrix are 16-dimensional vectors A_1, A_2, A_3 and A_4. We now discuss two cases:

- **Conventional Beamforming:** In order to do conventional beamforming the beamforming matrix
$$\mathcal{B} = \frac{(A^H A)^{-1} A^H}{\sqrt{Tr((A^H A)^{-1})}}$$
has to be computed. This matrix is then used for transmission.

- **Complementary Beamforming:** In addition to the above computation, we also need to compute an orthonormal basis of 16 dimensional vectors $U_0, U_1, U_2, \cdots, U_{11}$ for the orthogonal complement of the subspace spanned by the columns of A. This can be done using the Gram-Schmidt method and requires roughly the same number of operations as that required for the computation of $\mathcal{B}$. The matrices $Z_1, Z_2, \cdots, Z_L$ are constructed as below. When L is odd, we let $Z_L = 0$. For any L (either even or odd), we let $Z_{2i} = -Z_{2i-1}$ for $i = 1, 2, \cdots, \lfloor L/2 \rfloor$. The matrix Z_1, Z_3 and Z_5 are defined to have respectively rows equal to $U_0^H, U_1^H, U_2^H, U_3^H$, $U_4^H, U_5^H, U_6^H, U_7^H$, and $U_8^H, U_9^H, U_{10}^H, U_{11}^H$. We then periodically define

$$Z_1 = Z_7 = Z_{13} = \cdots,$$
$$Z_3 = Z_9 = Z_{15} = \cdots,$$
$$Z_5 = Z_{11} = Z_{17} = \cdots.$$

  Finally, we let

$$S^t = [((A^H A)^{-1} A^H / \sqrt{Tr((A^H A)^{-1})} + \frac{1}{\sqrt{k}} \epsilon Z_t]$$

  to be the complementary beamforming matrix at time t.

We thus observe from the above that complementary beamforming is approximately twice as much computationally intensive as conventional beamforming.

Remark on Network Impact: Soon after the invention of complementary beamforming, the network impact of the proposed method was considered in [6]. It was observed that using complementary beamforming significantly enhance the performance of a heavily loaded smart antenna enhanced IEEE 802.11 system as compared to the case when conventional beamforming is employed.

6 Conclusion

When a conventional beamformer is employed to increase the transmitted power to some specific directions, the radiated power to other directions is reduced. In a CSMA system, this implies that certain users may be nulled out by the beamformer under conventional beamforming. During a busy period of the channel, these users may wrongly determine that the channel is idle and start transmitting packets. This problem is referred to as the *hidden beam problem* which may cause unnecessary transmissions, subsequent back-offs, increased network latency and interference. Furthermore, the aforementioned undesired packet transmission has an energy penalty which adversely

effects the battery life of the remote devices. The hidden beam problem causes more severe degradation if the system is more heavily loaded which will be most likely the case both in hot spots and also whenever the system range is increased.

In this paper, in addressing the above problem, we proposed *complementary beamforming*, a new method of beamforming for wireless communications. In complementary beamforming, the objective function is both providing a minimum signal power to the silent users and increasing the signal power to the desired users. Using complementary beamforming, for smart antennas enhancements to CSMA systems (such as those of the IEEE 802.11), during a downlink busy period, other users can detect channel activity and remain silent.

We analyzed the proposed scheme for both the intended and silent users and proved that, when compared to conventional beamforming methods, for a negligible incurred power loss for the intended users, the effects of the hidden beam problem caused by the unintended users in the system can be significantly reduced.

Finally, we illustrated the proposed technique by some examples. We indicated that the complexity of complementary beamforming is approximately twice as much as that of the conventional beamforming.

7 Acknowledgment

We would like to thank Jim Brennan, Bobby Jose, Ed Casas, Tong Chia, Skip Crilly, Marcus da Silva, Praveen Mehrotra, and Hujun Yin for many insightful discussions.

References

[1] S. M. Alamouti, Y.-S. Choi and V. Tarokh, "Single-Beam Complementary Beamforming", *in preparation*, 2003.

[2] J. Bellorado, S.S. Ghassemzadeh, L.J. Greenstein, T. Sveinsson and V. Tarokh, "Coexistence of Ultra-Wideband Systems with IEEE-802.11a Wireless LANs", *Proceedings of VTC 2003*, to appear in Oct. 2003.

[3] Y.S. Choi, V. Tarokh and S.M. Alamouti, "A Subspace Approach to Complementary Beam-Forming", in preparation.

[4] Thomas M. Cover and Joy A Thomas, *Elements of Information Theory*, John Wiley and Sons, New York, 1991.

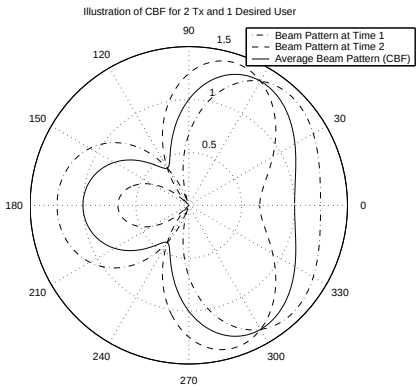


Figure 1: Illustration of Time Domain CBF pointing to $\theta = \frac{\pi}{3}$ and $d = \lambda/2$

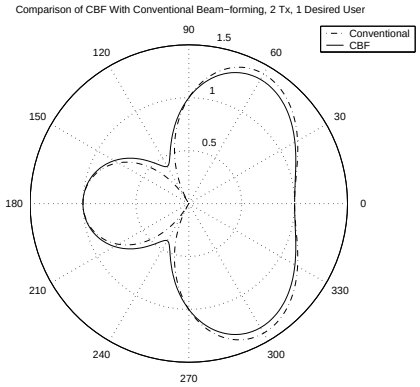


Figure 2: Comparison of CBF and Conventional Beamforming for $\epsilon = 0.3$.

[5] S.S. Ghassemzadeh, L.J. Greenstein, T. Sveinsson and V. Tarokh, "An Impulse Response Model For Residential Wireless Channels", *Proceedings of VTC 2003*, to appear in Oct. 2003.

[6] Praveen Mehrotra, Bobby Jose, Jim Brennan and Ed Casas, "Performance Impact of Smart Antennas on 802.11 MAC layer", *Proceedings of VTC 2003*, to appear in Oct. 2003.

[7] John G. Proakis, *Digital Communications*, Fourth Edition, McGraw-Hill Publishers, New York, 2000.

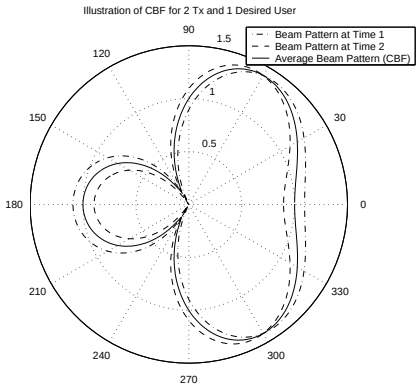


Figure 3: Illustration of Time Domain CBF pointing to $\theta = \frac{\pi}{3}$ and $d = \lambda/2$

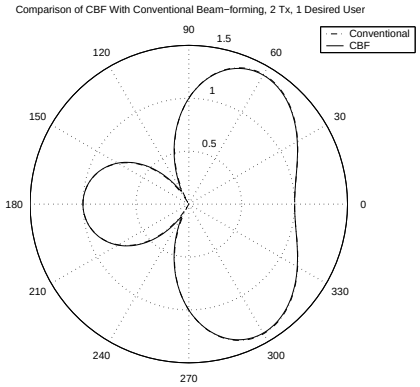


Figure 4: Comparison of CBF and Conventional Beamforming for $\epsilon = 0.1$.

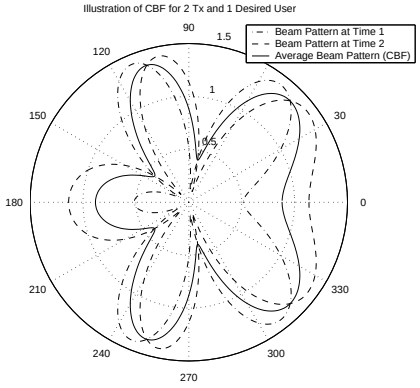


Figure 5: Illustration of Time Domain CBF pointing to $\theta = \frac{\pi}{4}$ and $d = \lambda$

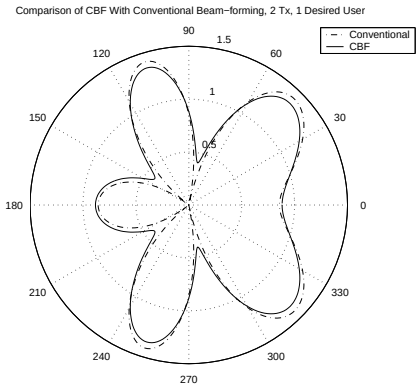


Figure 6: Comparison of CBF and Conventional Beamforming for $\epsilon = 0.3$.

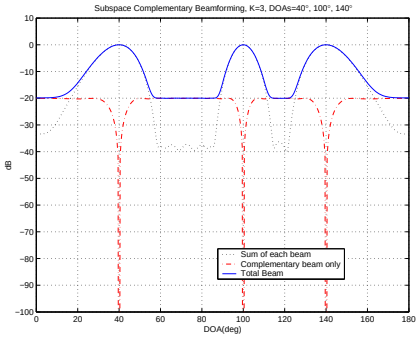


Figure 7: Beam Pattern For CBF

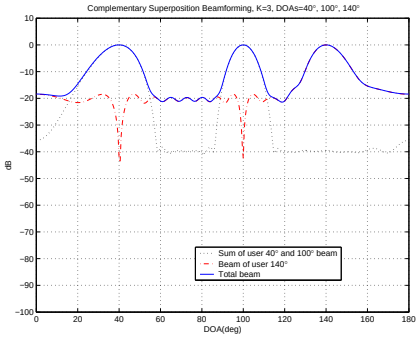


Figure 8: Beam Pattern for CBF